Bandits for BMO Functions

Tianyu Wang [1] **Cynthia Rudin** [1]

Abstract

We study the bandit problem where the underlying expected reward is a Bounded Mean Oscillation (BMO) function. BMO functions are allowed to be discontinuous and unbounded, and are useful in modeling signals with infinities in the domain. We develop a toolset for BMO bandits, and provide an algorithm that can achieve poly-log δ-regret – a regret measured against an arm that is optimal after removing a δ-sized portion of the arm space.

1. Introduction

Multi-Armed Bandit (MAB) problems model sequential decision making under uncertainty. Algorithms for this problem have important real-world applications including medical trials (Robbins, 1952) and web recommender systems (Li et al., 2010). While bandit methods have been developed for various settings, one problem setting that has not been studied, to the best of our knowledge, is when the expected reward function is a Bounded Mean Oscillation (BMO) function in a metric measure space. Intuitively, a BMO function does not deviate too much from its mean over any ball, and can be discontinuous or unbounded.

Such unbounded functions can model many real-world quantities. Consider the situation in which we are optimizing the parameters of a process (e.g., a physical or biological system) whose behavior can be simulated. The simulator is computationally expensive to run, which is why we could not exhaustively search the (continuous) parameter space for the optimal parameters. The "reward" of the system is sensitive to parameter values and can increase very quickly as the parameters change. In this case, by failing to model the infinities, even state-of-the-art continuum-armed bandit methods fail to compute valid confidence bounds, potentially leading to underexploration of the important part of the parameter space, and they may completely miss the optima.

As another example, when we try to determine failure modes of a system or simulation, we might try to locate singularities in the variance of its outputs. These are cases where the variance of outputs becomes extremely large. In this case, we can use a bandit algorithm for BMO functions to efficiently find where the system is most unstable.

There are several difficulties in handling BMO rewards. First and foremost, due to unboundedness in the *expected* reward functions, traditional regret metrics are doomed to fail. To handle this, we define a new performance measure, called δ-regret. The δ-regret measures regret against an arm that is optimal after removing a δ-sized portion of the arm space. Under this performance measure, and because the reward is a BMO function, our attention is restricted to a subspace on which the expected reward is finite. Subsequently, strategies that conform to the δ-regret are needed.

To develop a strategy that handles δ-regret, we leverage the John-Nirenberg inequality, which plays a crucial role in harmonic analysis. We construct our arm index using the John-Nirenberg inequality, in addition to a traditional UCB index. In each round, we play an arm with highest index. As we play more and more arms, we focus our attention on regions that contain good arms. To do this, we discretize the arm space adaptively, and carefully control how the index evolves with the discretization. We provide two algorithms – *Bandit-BMO-P* and *Bandit-BMO-Z*. They discretize the arm space in different ways. In *Bandit-BMO-P*, we keep a strict partitioning of the arm space. In *Bandit-BMO-Z*, we keep a collection of cubes where a subset of cubes form a discretization. *Bandit-BMO-Z* achieves poly-log δ-regret with high probability.

2. Related Works

Bandit problems in different settings have been actively studied since as far back as Thompson (1933). *Upper confidence bound* (UCB) algorithms remain popular (Robbins, 1952; Lai and Robbins, 1985; Auer, 2002) among the many approaches for (stochastic) bandit problems (e.g., see Srinivas et al., 2010; Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2012; Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014). Various extensions of upper confidence

[1]Department of Computer Science, Duke University, Durham, NC, USA. Correspondence to: Tianyu Wang <tianyu@cs.duke.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

bound algorithms have been studied. Some works use KL-divergence to construct the confidence bound (Lai and Robbins, 1985; Garivier and Cappé, 2011; Maillard et al., 2011), and some works include variance estimates within the confidence bound (Audibert et al., 2009; Auer and Ortner, 2010). UCB is also used in the contextual setting (e.g., Li et al., 2010; Krause and Ong, 2011; Slivkins, 2014).

Perhaps Lipschitz bandits are closest to BMO bandits. The Lipschitz bandit problem was termed "continuum-armed bandits" in early stages (Agrawal, 1995). In "continuum-armed bandits," arm space is continuous – e.g., $[0, 1]$. Along this line, bandits that are Lipschitz continuous (or Hölder continuous) have been studied. In particular, Kleinberg (2005) proves a $\Omega(T^{2/3})$ lower bound and proposes a $\widetilde{\mathcal{O}}\left(T^{2/3}\right)$ algorithm. Under other extra conditions on top of Lipschitzness, regret rate of $\widetilde{\mathcal{O}}(T^{1/2})$ was achieved (Cope, 2009; Auer et al., 2007). For general (doubling) metric spaces, the Zooming bandit algorithm (Kleinberg et al., 2008) and Hierarchical Optimistic Optimization algorithm (Bubeck et al., 2011a) were developed. In more recent years, some attention has been given to Lipschitz bandit problems with certain extra conditions. To name a few, Bubeck et al. (2011b) study Lipschitz bandits for differentiable rewards, which enables algorithms to run without explicitly knowing the Lipschitz constants. The idea of robust mean estimators (Bubeck et al., 2013; Bickel et al., 1965; Alon et al., 1999) was applied to the Lipschitz bandit problem to cope with heavy-tail rewards, leading to the development of a near-optimal algorithm (Lu et al., 2019). Lipschitz bandits with an unknown metric, where a clustering is used to infer the underlying unknown metric, has been studied by Wanigasekara and Yu (2019). Lipschitz bandits with discontinuous but bounded rewards were studied by Krishnamurthy et al. (2019).

An important setting that is beyond the scope of the aforementioned works is when the expected reward is allowed to be unbounded. This setting breaks the previous Lipschitzness assumption or "almost Lipschitzness" assumption (Krishnamurthy et al., 2019), which may allow discontinuities but require boundedness. To the best of our knowledge, this paper is the first work that studies the bandit learning problem for BMO functions.

3. Preliminaries

We review the concept of (rectangular) Bounded Mean Oscillation (BMO) in Euclidean space (e.g., Fefferman, 1979; Stein and Murphy, 1993).

Definition 1. *(BMO Functions) Let $(\mathbb{R}^d, \mu)$ be the Euclidean space with the Lebesgue measure. Let $L^1_{loc}(\mathbb{R}^d, \mu)$ denote the space of measurable functions (on $\mathbb{R}^d$) that are locally integrable with respect to μ. A function $f \in L^1_{loc}(\mathbb{R}^d, \mu)$ is said to be a Bounded Mean Oscillation function, $f \in BMO(\mathbb{R}^d, \mu)$, if there exists a constant C_f, such that for any hyper-rectangles $Q \subset \mathbb{R}^d$,*

$$\frac{1}{\mu(Q)} \int_Q |f - \langle f \rangle_Q | d\mu \leq C_f, \quad \langle f \rangle_Q := \frac{\int_Q f d\mu}{\mu(Q)}. \quad (1)$$

For a given such function f, the infimum of the admissible constant C_f over all hyper-rectangles Q is denoted by $\|f\|_{BMO}$, or simply $\|f\|$. We use $\|f\|_{BMO}$ and $\|f\|$ interchangeably in this paper.

A BMO function can be discontinuous and unbounded. The function in Figure 1 illustrates the singularities a BMO function can have over its domain. Our problem is most interesting when multiple singularities of this kind occur.

To properly handle the singularities, we will need the John-Nirenberg inequality (Theorem 1), which plays a central role in our paper.

Theorem 1 (John-Nirenberg inequality)**.** *Let μ be the Lebesgue measure. Let $f \in BMO\left(\mathbb{R}^d, \mu\right)$. Then there exists constants C_1 and C_2, such that, for any hypercube $q \subset \mathbb{R}^d$ and any $\lambda > 0$,*

$$\mu\left(\left\{x \in q : \left|f(x) - \langle f \rangle_q\right| > \lambda\right\}\right) \leq C_1 \mu(q) \exp\left\{\frac{-\lambda}{C_2 \|f\|}\right\}. \quad (2)$$

The John-Nirenberg inequality dates back to at least John (1961), and a proof is provided in Appendix C.

As shown in Appendix C, $C_1 = e$ and $C_2 = e2^d$ provide a pair of legitimate C_1, C_2 values. However, this pair of C_1 and C_2 values may be overly conservative. Tight values of C_1 and C_2 are not known in general cases (Lerner, 2013; Slavin and Vasyunin, 2017), and it is also conjectured that C_2 and C_1 might be independent of dimension (Cwikel et al., 2012). For the rest of the paper, we use $\|f\| = 1$, $C_1 = 1$, and $C_2 = 1$, which permits cleaner proofs. Our results generalize to cases where C_1, C_2 and $\|f\|$ are other constant values.

In this paper, we will work in Euclidean space with the Lebesgue measure. For our purpose, Euclidean space is as general as doubling spaces, since we can always embed a doubling space into a Euclidean space with some distortion of metric. This fact is formally stated in Theorem 2.

Theorem 2. *(Assouad, 1983). Let (X, d) be a doubling metric space and $\varsigma \in (0, 1)$. Then (X, d^ς) admits a bi-Lipschitz embedding into $\mathbb{R}^n$ for some $n \in \mathbb{N}$.*

In a doubling space, any ball of radius ρ can be covered by M_d balls of radius $\frac{\rho}{2}$, where M_d is the ***doubling constant***. In the space $(\mathbb{R}^d, \|\cdot\|_\infty)$, the doubling constant M_d is 2^d. In domains of other geometries, the doubling constant can

be much smaller than exponential. Throughout the rest of the paper, we use M_d to denote the doubling constant.

4. Problem Setting: BMO Bandits

The goal of a stochastic bandit algorithm is to exploit the current information, and explore the space efficiently. In this paper, we focus on the following setting: a payoff function is defined over the arm space $([0,1)^d, \|\cdot\|_{\max}, \mu)$, where μ is the Lebesgue measure (note that $[0,1)^d$ is a Lipschitz domain). The payoff function is:

$$f : [0,1)^d \to \mathbb{R} \quad \text{where} \quad f \in BMO([0,1)^d, \mu). \tag{3}$$

The actual observations are given by $y(a) = f(a) + \mathcal{E}_a$, where $\mathcal{E}_a$ is a zero-mean noise random variable whose distribution can change with a. We assume that for all a, $|\mathcal{E}_a| \le D_{\mathcal{E}}$ almost surely for some constant $D_{\mathcal{E}}$ (**N1**). Our results generalize to the setting with sub-Gaussian noise (Shamir, 2011). We also assume that the expected reward function does not depend on noise.

In our setting, an agent is interacting with this environment in the following fashion. At each round t, based on past observations $(a_1, y_1, \cdots, a_{t-1}, y_{t-1})$, the agent makes a query at point a_t and observes the (noisy) payoff y_t, where y_t is revealed only after the agent has made a decision a_t. For a payoff function f and an arm sequence $a_1, a_2, \cdots, a_T$, we use δ-regret incurred up to time T as the performance measure (Definition 2).

Definition 2. *(δ-regret) Let $f \in BMO([0,1)^d, \mu)$. A number $\delta \ge 0$ is called f-`admissible` if there exists a real number z_0 that satisfies*

$$\mu(\{a \in [0,1)^d : f(a) > z_0\}) = \delta. \tag{4}$$

For an f-admissible δ, define the set F^δ to be

$$F^\delta := \left\{z \in \mathbb{R} : \mu(\{a \in [0,1)^d : f(a) > z\}) = \delta\right\}. \tag{5}$$

Define $f^\delta := \inf F^\delta$. For a sequence of arms $A_1, A_2, \cdots$, and σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \cdots$ where $\mathcal{F}_t$ describes all randomness before arm A_t, define the δ-regret at time t as

$$r_t^\delta := \max\{0, f^\delta - \mathbb{E}_t[f(A_t)]\}, \tag{6}$$

where $\mathbb{E}_t$ is the expectation conditioned on $\mathcal{F}_t$. The total δ-regret up to time T is then $R_T^\delta := \sum_{t=1}^T r_t^\delta$.

Intuitively, the δ-regret is measured against an amended reward function that is created by chopping off a small portion of the arm space where the reward may become unbounded. As an example, Figure 1 plots a BMO function and its f^δ value. A problem defined as above with performance measured by δ-regret is called a ***BMO bandit problem***.

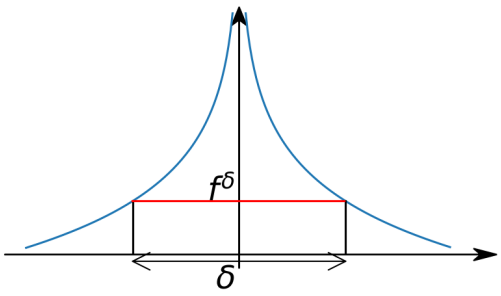

Figure 1. Graph of $f(x) = -\log(|x|)$, with δ and f^δ annotated. This function is an unbounded BMO function.

Remark 1. *The definition of δ-regret, or a definition of this kind, is needed for a reward function $f \in BMO([0,1)^d, \mu)$. For an unbounded BMO function f, the max value is infinity, while f^δ is a finite number as long as δ is f-admissible.*

Remark 2 (Connection to bandits with heavy-tails)**.** *In the definition of bandits with heavy tails (Bubeck et al., 2013; Medina and Yang, 2016; Shao et al., 2018; Lu et al., 2019), the reward distribution at a fixed arm is heavy-tail – having a bounded expectation and bounded $(1+\beta)$-moment ($\beta \in (0,1]$). In the case of BMO rewards, the expected reward itself can be unbounded. Figure 1 gives an instance of unbounded BMO reward, which means the BMO bandit problem is not covered by settings of bandits with heavy tails.*

A quick consequence of the definition of δ-regret is the following lemma. This lemma is used in the regret analysis when handling the concentration around good arms.

Lemma 1. *Let f be the reward function. For any f-admissible $\delta \ge 0$, let $S^\delta := \left\{a \in [0,1)^d : f(a) > f^\delta\right\}$. Then we have S^δ measurable and $\mu(S^\delta) = \delta$.*

Before moving on to the algorithms, we put forward the following assumption.

Assumption 1. *We assume that the expected reward function $f \in BMO([0,1)^d, \mu)$ satisfies $\langle f \rangle_{[0,1)^d} = 0$.*

Assumption 1 does not sacrifice generality. Since f is a BMO function, it is locally-integrable. Thus $\langle f \rangle_{[0,1)^d}$ is finite, and we can translate the reward function up or down such that $\langle f \rangle_{[0,1)^d} = 0$.

5. Solve BMO Bandits via Partitioning

BMO bandit problems can be solved by partitioning the arm space and treating the problem as a finite-arm problem among partitions. For our purpose, we maintain a sequence

of partitions using *dyadic cubes*. By *dyadic cubes* of $\mathbb{R}^d$, we refer to the collection of all cubes of the following form:

$$\mathcal{Q}_{\mathbb{R}^d} := \left\{\Pi_{i=1}^d \left[m_i 2^{-k}, m_i 2^{-k} + 2^{-k}\right)\right\} \tag{7}$$

where Π is the Cartesian product, and $m_1, \cdots, m_d, k \in \mathbb{Z}$. Dyadic cubes of $[0,1)^d$ is $\mathcal{Q}_{[0,1)^d} := \left\{q \in \mathcal{Q}_{\mathbb{R}^d} : q \subset [0,1)^d\right\}$. Dyadic cubes of $[0,1)^2$ are $\{[0,1)^2, [0,0.5)^2, [0.5,1)^2, [0.5,1) \times [0,0.5), \cdots\}$.

We say a dyadic cube Q is a ***direct sub-cube*** of a dyadic cube Q' if $Q \subseteq Q'$ and the edge length of Q' is twice the edge length of Q. By definition of doubling constant, for any cube Q, it has M_d direct sub-cubes, and these direct sub-cubes form a partition of Q. If Q is a direct sub-cube of Q', then Q' is a ***direct super cube*** of Q.

At each step t, Bandit-BMO-P treats the problem as a finite-arm bandit problem with respect to the cubes in the dyadic partition at t; each cube possesses a confidence bound. The algorithm then chooses a best cube according to UCB, and chooses an arm uniformly at random within the chosen cube. Before formulating our strategy, we put forward several functions that summarize cube statistics.

Let $\mathcal{Q}_t$ be the collection of dyadic cubes of $[0,1)^d$ at time t ($t \geq 1$). Let $(a_1, y_1, a_2, y_2, \cdots, a_t, y_t)$ be the observations received up to time t. We define

- the *cube count* $n_t : \mathcal{Q}_t \to \mathbb{R}$, such that for $q \in \mathcal{Q}_t$
$$n_t(q) := \sum_{i=1}^{t-1} \mathbb{I}_{[a_i \in q]}; \quad \widetilde{n}_t(q) := \max(1, n_t(q)). \tag{8}$$
- the *cube average* $m_t : \mathcal{Q}_t \to \mathbb{R}$, such that for $q \in \mathcal{Q}_t$
$$m_t(q) := \begin{cases} \frac{\sum_{i=1}^{t-1} y_i \mathbb{I}_{[a_i \in q]}}{n_t(q)}, & \text{if } n_t(q) > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

At time t, based on the partition $\mathcal{Q}_{t-1}$ and observations $(a_1, y_1, a_2, y_2, \cdots, a_{t-1}, y_{t-1})$, our bandit algorithm picks a cube (and plays an arm within the cube uniformly at random). More specifically, the algorithm picks

$$Q_t \in \arg\max_{q \in \mathcal{Q}_t} U_t(q), \quad \text{where} \tag{10}$$

$$U_t(q) := m_t(q) + H_t(q) + J(q), \tag{11}$$

$$H_t(q) := \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\widetilde{n}_t(q)}},$$

$$J(q) := \lfloor \log(\mu(q)/\eta) \rfloor_+ ,$$

$$\Psi := \max\left\{\log(T^2/\epsilon), 2\log_2(1/\eta)\right\}, \tag{12}$$

where T is the time horizon, $D_{\mathcal{E}}$ is the a.s. bound on the noise, ϵ and η are algorithm parameters (to be discussed in more detail later), and $\lfloor z \rfloor_+ = \max\{0, z\}$. Here Ψ is the "Effective Bound" of the expected reward, and η controls minimal cube size in the partition $\mathcal{Q}_t$ (Proposition 3 in Appendix A.4). All these quantities will be discussed in more detail as we develop our algorithm.

After playing an arm and observing reward, we update the partition into a finer one if needed. Next, we discuss our partition refinement rules and the tie-breaking mechanism.

Partition Refinement: We start with $\mathcal{Q}_0 = \{[0,1)^d\}$. At time t, we split cubes in $\mathcal{Q}_{t-1}$ to construct $\mathcal{Q}_t$ so that the following is satisfied for any $q \in \mathcal{Q}_t$

$$H_t(q) \geq J(q), \quad \text{or equivalently}$$
$$\frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\widetilde{n}_t(q)}} \geq \lfloor \log(\mu(q)/\eta) \rfloor_+ . \tag{13}$$

In (13), the left-hand-side does not decrease as we make splits (the numerator remains constant while the denominator can only decrease), while the right-hand-side decreases until it hits zero as we make more splits. Thus (13) can always be satisfied with additional splits.

Tie-breaking: We break down our tie-breaking mechanism into two steps. In the first step, we choose a cube $Q_t \in \mathcal{Q}_{t-1}$ such that:

$$Q_t \in \arg\max_{q \in \mathcal{Q}_{t-1}} U_t(q). \tag{14}$$

After deciding from which cube to choose an arm, we uniformly randomly play an arm A_t within the cube Q_t. If measure μ is non-uniform, we play arm A_t, so that for any subset $S \subset Q_t$, $\mathbb{P}(A_t \in S) = \frac{\mu(S)}{\mu(Q_t)}$.

The random variables $\{(Q_{t'}, A_{t'}, Y_{t'})\}_{t'}$ (*cube selection*, *arm selection*, *reward*) describe all randomness in the learning process up to time t. We summarize this strategy in Algorithm 1. Analysis of Algorithm 1 is found in Section 5.1, which also provides some tools for handling δ-regret. Then in Section 6, we provide an improved algorithm that exhibits a stronger performance guarantee.

5.1. Regret Analysis of Bandit-BMO-P

In this section we provide a theoretical guarantee on the algorithm. We will use capital letters (e.g., Q_t, A_t, Y_t) to denote random variables, and use lower-case letters (e.g. a, q) to denote non-random quantities, unless otherwise stated.

Theorem 3. *Fix any T. With probability at least $1 - 2\epsilon$, for any $\delta > |\mathcal{Q}_T|\eta$ such that δ is f-admissible, the total δ-regret for Algorithm 1 up to time T satisfies*

$$\sum_{t=1}^{T} r_t^{\delta} \lesssim_d \widetilde{\mathcal{O}}\left(\sqrt{T|\mathcal{Q}_T|}\right), \tag{15}$$

where the $\lesssim_d$ sign omits constants that depends on d, and $|\mathcal{Q}_T|$ is the cardinality of $\mathcal{Q}_T$.

Algorithm 1 Bandit-BMO-Partition (Bandit-BMO-P)

Problem intrinsics: $\mu(\cdot)$, $D_{\mathcal{E}}$, d, M_d.
$\{\mu(\cdot)$ is the Lebesgue measure. $D_{\mathcal{E}}$ bounds the noise.$\}$
$\{d$ is the dimension of the arm space.$\}$
$\{M_d$ is the doubling constant of the arm space.$\}$
Algorithm parameters: $\eta > 0$, $\epsilon > 0$, T.
$\{T$ is the time horizon. ϵ and η are parameters.$\}$
for $t = 1, 2, \ldots, T$ **do**
  Let m_t and n_t be defined as in (9) and (8).
  Select a cube $Q_t \in \mathcal{Q}_t$ such that:

$$Q_t \in \arg\max_{q \in \mathcal{Q}_{t-1}} U_t(l),$$

  where U_t is defined in (11).
  Play arm $A_t \in Q_t$ uniformly at random. Observe Y_t.
  Update the partition $\mathcal{Q}_t$ to $\mathcal{Q}_{t+1}$ according to (13).
end for

From uniform tie-breaking, we have

$$\mathbb{E}[f(A_t)|\mathcal{F}_t] = \frac{1}{\mu(Q_t)} \int_{a \in Q_t} f(a)\, da = \langle f \rangle_{Q_t}, \quad (16)$$

$$\mathcal{F}_t = \sigma(Q_1, A_1, Y_1, \cdots, Q_{t-1}, A_{t-1}, Y_{t-1}, Q_t), \quad (17)$$

where $\mathcal{F}_t$ is the σ-algebra generated by random variables $Q_1, A_1, Y_1, \cdots, Q_{t-1}, A_{t-1}, Y_{t-1}, Q_t$ – all randomness right after selecting cube Q_t. At time t, the expected reward is the mean function value of the selected cube.

The proof of the theorem is divided into two parts. In **Part I**, we show that some "good event" holds with high probability. In **Part II**, we bound the δ-regret under the "good event."

Part I: For $t \le T$, and $q \in \mathcal{Q}_t$, we define

$$\mathcal{E}_t(q) := \left\{ \left| \langle f \rangle_q - m_t(q) \right| \le H_t(q) \right\}, \quad (18)$$

$$H_t(q) = \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\widetilde{n}_t(q)}}. \quad (19)$$

In the above, $\mathcal{E}_t(q)$ is essentially saying that the empirical mean within a cube q concentrates to $\langle f \rangle_q$. Lemma 2 shows that $\mathcal{E}_t(q)$ happens with high probability for any t and q.

Lemma 2. *With probability at least $1 - \frac{\epsilon}{T}$, the event $\mathcal{E}_t(q)$ holds for any $q \in \mathcal{Q}_t$ at any time t.*

To prove Lemma 2, we apply a variation of Azuma's inequality (Vu, 2002; Tao and Vu, 2015). We also need some additional effort to handle the case when a cube q contains no observations. The details are in Appendix A.3.

Part II: Next, we link the δ-regret to the $J(q)$ term.

Lemma 3. *Recall $J(q) = \log(\mu(q)/\eta)$. For any partition $\mathcal{Q}$ of $[0,1)^d$, there exists $q \in \mathcal{Q}$, such that*

$$f^\delta - \langle f \rangle_q \le J(q), \quad (20)$$

for any f-admissible $\delta > \eta|\mathcal{Q}|$, where $|\mathcal{Q}|$ is the cardinality of $\mathcal{Q}$.

In the proof of Lemma 3, we suppose, in order to get a contradiction, that there is no such cube. Under this assumption, there will be contradiction to the definition of f^δ.

By Lemma 3, there exists a "good" cube $\widetilde{q}_t$ (at any time $t \le T$), such that (20) is true for $\widetilde{q}_t$. Let δ be an arbitrary number satisfying (1) $\delta > |\mathcal{Q}_T|\eta$ and (2) δ is f-admissible. Then under event $\mathcal{E}(\widetilde{q}_t)$,

$$\begin{aligned} f^\delta &= \left(f^\delta - \langle f \rangle_{\widetilde{q}_t}\right) + \left(\langle f \rangle_{\widetilde{q}_t} - m_t(\widetilde{q}_t)\right) + m_t(\widetilde{q}_t) \\ &\overset{①}{\le} J(\widetilde{q}_t) + H_t(\widetilde{q}_t) + m_t(\widetilde{q}_t), \end{aligned} \quad (21)$$

where ① uses Lemma 3 for the first brackets and Lemma 2 (with event $\mathcal{E}_t(\widetilde{q}_t)$) for the second brackets.

The event where all "good" cubes and all cubes we select (for $t \le T$) have nice estimates, namely $\left(\bigcap_{t=1}^T \mathcal{E}_t(\widetilde{q}_t)\right) \bigcap \left(\bigcap_{t=1}^T \mathcal{E}_t(Q_t)\right)$, occurs with probability at least $1 - 2\epsilon$. This result comes from Lemma 2 and a union bound, and we note that $\mathcal{E}_t(q)$ depends on ϵ (and T), as in (19). Under this event, from (18) we have $\left|\langle f \rangle_{Q_t} - m_t(Q_t)\right| \le H_t(Q_t)$. This and (16) give us

$$\mathbb{E}\left[f(A_t)|\mathcal{F}_t\right] = \langle f \rangle_{Q_t} \ge m_t(Q_t) - H_t(Q_t). \quad (22)$$

We can then use the above to get, under the "good event",

$$\begin{aligned} & f^\delta - \mathbb{E}[f(A_t)|\mathcal{F}_t] \\ \overset{①}{\le}& m_t(\widetilde{q}_t) + H_t(\widetilde{q}_t) + J(\widetilde{q}_t) - m_t(Q_t) + H_t(Q_t) \\ \overset{②}{\le}& m_t(Q_t) + H_t(Q_t) + J(Q_t) - m_t(Q_t) + H_t(Q_t) \\ =& 2H_t(Q_t) + J(Q_t) \le 3H_t(Q_t), \end{aligned} \quad (23)$$

where ① uses (21) for the first three terms and (22) for the last three terms, ② uses that $U_t(Q_t) \ge U_t(\widetilde{q}_t)$ since Q_t maximizes the index $U_t(\cdot)$ according to (14), and the last inequality uses the rule (13).

Next, we use Lemma 4 to link the number of cubes up to a time t to the Hoeffding-type tail bound in (23). Intuitively, this bound (Lemma 4) states that the numbers of points within the cubes grows fast enough to be bounded by a function of the number of cubes.

Lemma 4. *We say a partition $\mathcal{Q}$ is finer than a partition $\mathcal{Q}'$ if for any $q \in \mathcal{Q}$, there exists $q' \in \mathcal{Q}'$ such that $q \subset q'$. Consider an arbitrary sequence of points $x_1, x_2, \cdots, x_t, \cdots$ in a space $\mathcal{X}$, and a sequence of partitions $\mathcal{Q}_1, \mathcal{Q}_2, \cdots$ of $\mathcal{X}$ such that $\mathcal{Q}_{t+1}$ is finer than $\mathcal{Q}_t$ for all $t = 1, 2, \cdots, T-1$. Then for any T, and $\{q_t \in \mathcal{Q}_t\}_{t=1}^T$,*

$$\sum_{t=1}^T \frac{1}{\widetilde{n}_t(q_t)} \le e|\mathcal{Q}_T| \log\left(1 + (e-1)\frac{T}{|\mathcal{Q}_T|}\right), \quad (24)$$

where $\widetilde{n}_t$ is defined in (8) (using points $x_1, x_2, \cdots$), and $|\mathcal{Q}_t|$ is the cardinality of partition $\mathcal{Q}_t$.

A proof of Lemma 4 is in Appendix B. We can apply Lemma 4 and the Cauchy-Schwarz inequality to (23) to prove Theorem 3. The details can be found in Appendix A.7.

6. Achieve Poly-log Regret via Zooming

In this section we study an improved version of the previous section that uses the Zooming machinery (Kleinberg et al., 2008; Slivkins, 2014) and inspirations from Bubeck et al. (2011a). Similar to Algorithm 1, this algorithm runs by maintaining a set of dyadic cubes $\mathcal{Q}_t$.

In this setting, we divide the time horizon into episodes. In each episode t, we are allowed to play multiple arms, and all arms played can incur regret. This is also a UCB strategy, and the index of $q \in \mathcal{Q}_t$ is defined the same way as (11):

$$U_t(q) := m_t(q) + H_t(q) + J(q) \tag{25}$$

Before we discuss in more detail how to select cubes and arms based on the above index $U_t(\cdot)$, we first describe how we maintain the collection of cubes. Let $\mathcal{Q}_t$ be the collection of dyadic cubes at episode t. We first define ***terminal cubes***, which are cubes that do not have sub-cubes in $\mathcal{Q}_t$. More formally, a cube $Q \in \mathcal{Q}_t$ is a terminal cube if there is no other cube $Q' \in \mathcal{Q}_t$ such that $Q' \subset Q$. A ***pre-parent cube*** is a cube in $\mathcal{Q}_t$ that "directly" contains a terminal cube: For a cube $Q \in \mathcal{Q}_t$, if Q is a direct super cube of any terminal cube, we say Q is a pre-parent cube. Finally, for a cube $Q \in \mathcal{Q}_t$, if Q is a pre-parent cube and no super cube of Q is a pre-parent cube, we call Q a ***parent cube***. Intuitively, no "sibling" cube of a parent cube is a terminal cube. As a consequence of this definition, a parent cube cannot contain another parent cube. Note that some cubes are none of the these three types of cubes. Figure 2 gives examples of terminal cubes, pre-parent cubes and parent cubes.

Algorithm Description

Pick ***zooming rate*** $\alpha \in \left(0, \frac{(\Psi+D_{\mathcal{I}})\sqrt{2\log(2T^2/\epsilon)}}{\log(M_d/\eta)}\right]$. The collection of cubes grows following the rules below: **(1)** Initialize $\mathcal{Q}_0 = \{[0,1)^d\}$ and $[0,1)^d$. Warm-up: play n_{warm} arms uniformly at random from $[0,1)^d$ so that

$$\begin{cases} \frac{(\Psi+D_{\mathcal{I}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{n_{warm}}} \geq \alpha \log\left(\frac{M_d}{\eta}\right) \\ \frac{(\Psi+D_{\mathcal{I}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{n_{warm}+1}} < \alpha \log\left(\frac{M_d}{\eta}\right) \end{cases}. \tag{26}$$

(2) After episode t ($t = 1, 2, \cdots, T$), ensure

$$\frac{(\Psi+D_{\mathcal{I}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\widetilde{n}_t(Q^{ter})}} \geq \alpha \log\left(\frac{M_d \mu(Q^{ter})}{\eta}\right) \tag{27}$$

for any terminal cube Q^{ter}. If (27) is violated for a terminal cube Q^{ter}, we include the M_d direct sub-cubes of Q^{ter} into $\mathcal{Q}_t$. Then Q^{ter} will no longer be a terminal cube and the direct sub-cubes of Q^{ter} will be terminal cubes. We repeatedly include direct sub-cubes of (what were) terminal cubes into $\mathcal{Q}_t$, until all terminal cubes satisfy (27). We choose α to be smaller than $\frac{(\Psi+D_{\mathcal{I}})\sqrt{2\log(2T^2/\epsilon)}}{\log(M_d/\eta)}$ so that (27) can be satisfied with $\widetilde{n}_t(Q^{ter}) = 1$ and $\mu(Q^{ter}) = 1$.

As a consequence, any non-terminal cube Q^{par} (regardless of whether it is a pre-parent or parent cube) satisfies:

$$\frac{(\Psi+D_{\mathcal{I}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\widetilde{n}_t(Q^{par})}} < \alpha \log\left(\frac{M_d \mu(Q^{par})}{\eta}\right). \tag{28}$$

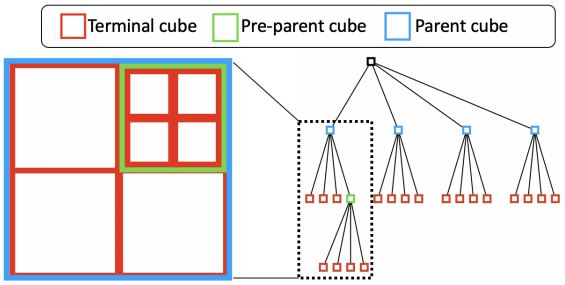

Figure 2. Example of terminal cubes, pre-parent and parent cubes.

After the splitting rule is achieved, we select a parent cube. Specifically Q_t is chosen to maximize the following index:

$$Q_t \in \arg \max_{q \in \mathcal{Q}_t, q \text{ is a parent cube}} U_t(q).$$

Within each *direct sub-cube* of Q_t (either pre-parent or terminal cubes), we uniformly randomly play one arm. In each episode t, M_d arms are played. This algorithm is summarized in Algorithm 2.

Regret Analysis: For the rest of the paper, we define

$$\mathcal{F}_t := \sigma\left(\left\{Q_{t'}, \{A_{t',j}\}_{j=1}^{M_d}, \{Y_{t',j}\}_{j=1}^{M_d}\right\}_{t'=1}^{t-1}, Q_t\right),$$

which is the σ-algebra describing all randomness right after selecting the parent cube for episode t. We use $\mathbb{E}_t$ to denote the expectation conditioning on $\mathcal{F}_t$. We will show Algorithm 2 achieves $\widetilde{\mathcal{O}}\left(poly\text{-}log(T)\right)$ δ-regret with high probability (formally stated in Theorem 4).

Let $A_{t,i}$ be the i-th arm played in episode t. Let us denote $\Delta_{t,i}^{\delta} := f^{\delta} - \mathbb{E}_t[f(A_{t,i})]$. Since each $A_{t,i}$ is selected uniformly randomly within a direct sub-cube of Q_t, we have

$$\sum_{i=1}^{M_d} \mathbb{E}_t[f(A_{t,i})] = M_d \langle f \rangle_{Q_t}, \tag{29}$$

Algorithm 2 Bandit-BMO-Zooming (Bandit-BMO-Z)

Problem intrinsics: $\mu(\cdot)$, $D_{\mathbb{E}}$, d, M_d.
$\{\mu(\cdot), D_{\mathbb{E}}, d, M_d$ are same as those in Algorithm 1.$\}$
Algorithm parameters: $\eta, \epsilon, T > 0$, and $\alpha \in \left(0, \frac{(\Psi+D_{\mathbb{E}})\sqrt{2\log(2T^2/\epsilon)}}{\log(M_d/\eta)}\right]$.
$\{\eta$, ϵ, T are same as those in Algorithm 1. α is the zooming rate.$\}$
Initialize: let $\mathcal{Q}_0 = [0,1)^d$. Play warm-up phase (26).
for episode $t = 1, 2, \ldots, T$ **do**
  Let m_t, n_t, U_t be defined as in (9), (8) and (25).
  Select parent cube $Q_t \in \mathcal{Q}_t$ such that:

$$Q_t \in \arg \max_{q \in \mathcal{Q}_t,\ q \text{ is a parent cube.}} U_t(q).$$

  for $j = 1, 2, \ldots, M_d$ **do**
    Locate the j-th direct sub-cube of Q_t: Q_j^{sub}.
    Play $A_{t,j} \in Q_j^{sub}$ uniformly at random, and observe $Y_{t,j}$.
  end for
  Update the collection of dyadic cubes $\mathcal{Q}_t$ to $\mathcal{Q}_{t+1}$ according to (27).
end for

where $\mathbb{E}_t$ is the expectation conditioning on all randomness before episode t. Using the above equation, for any t,

$$\sum_{i=1}^{M_d} \Delta_{t,i}^{\delta} = M_d(f^{\delta} - \langle f \rangle_{Q_t}). \tag{30}$$

The quantity $\sum_{i=1}^{M_d} \Delta_{t,i}^{\delta}$ is the δ-regret incurred during episode t. We will bound (30) using tools in Section 5. In order to apply Lemma 3, we need to show that the parent cubes form of partition of the arm space (Proposition 1).

Proposition 1. *At any episode t, the collection of parent cubes forms a partition of the arm space.*

Since the parent cubes in $\mathcal{Q}_t$ form a partition of the arm space, we can apply Lemma 3 to get the following. For any episode t, there exists a parent cube $q_t^{\max}$, such that

$$f^{\delta} \le \langle f \rangle_{q_t^{\max}} + \log\left(\mu(q_t^{\max})/\eta\right). \tag{31}$$

Let us define $\widetilde{\mathcal{E}}_T := \left(\bigcap_{t=1}^{T} \mathcal{E}_t(q_t^{\max})\right) \bigcap \left(\bigcap_{t=1}^{T} \mathcal{E}_t(Q_t)\right)$, where $\mathcal{E}_t(q_t^{\max})$ and $\mathcal{E}_t(Q_t)$ are defined in (18). By Lemma 2 and another union bound, we know the event $\widetilde{\mathcal{E}}_T$ happens with probability at least $1 - 2\epsilon$.

Since each episode creates at most a constant number of new cubes, we have $|\mathcal{Q}_t| = \mathcal{O}(t)$. Using the argument we used for (23), we have that at any $t \le T$, for any $\delta > \eta|\mathcal{Q}_t|$ that is f-admissible, under event $\widetilde{\mathcal{E}}_T$,

$$\sum_{i=1}^{M_d} \Delta_{t,i}^{\delta} = M_d\left(f^{\delta} - \langle f \rangle_{Q_t}\right) \tag{32}$$

$$\le M_d\left(2\frac{(\Psi+D_{\mathbb{E}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\widetilde{n}_t(Q_t)}} + \log\left(\frac{M_d\mu(Q_t)}{\eta}\right)\right)$$

$$\le M_d(1+2\alpha)\log\left(\frac{M_d\mu(Q_t)}{\eta}\right), \tag{33}$$

where (32) uses (30) and the last inequality uses (28).

Next, we extend some definitions from Kleinberg et al. (2008), to handle the δ-regret setting. Firstly, we define the set of (λ, δ)-optimal arms as

$$\mathcal{X}_{\delta}(\lambda) := \left(\bigcup\{Q \subset [0,1]^d : f^{\delta} - \langle f \rangle_Q \le \lambda\}\right). \tag{34}$$

We also need to extend the definition of zooming number (Kleinberg et al., 2008) to our setting. We denote by $N_{\delta}(\lambda, \xi)$ the number of cubes of edge-length ξ needed to cover the set $\mathcal{X}_{\delta}(\lambda)$. Then we define the (δ, η)-Zooming Number with zooming rate α as

$$\widetilde{N}_{\delta,\eta,\alpha} := \sup_{\lambda \in \left(\eta^{\frac{1}{d}}, 1\right]} N_{\delta}\left((1+2\alpha)\log\left(M_d\lambda^d/\eta\right), \lambda\right), \tag{35}$$

where $N_{\delta}\left((1+2\alpha)\log\left(M_d\lambda^d/\eta\right), \lambda\right)$ is the number of cubes of edge-length λ needed to cover $\mathcal{X}_{\delta}((1+2\alpha)\log(M_d\lambda^d/\eta))$. The number $\widetilde{N}_{\delta,\eta,\alpha}$ is well-defined. This is because the $\mathcal{X}_{\delta}((1+2\alpha)\log(M_d\lambda^d/\eta))$ is a subspace of $(0,1]^d$, and number of cubes of edge-length $> \eta^{\frac{1}{d}}$ needed to cover $(0,1]^d$ is finite. Intuitively, the idea of zooming is to use smaller cubes to cover more optimal arms, and vice versa. BMO properties convert between units of reward function and units in arm space.

We will regroup the $\Delta_{t,i}$ terms to bound the regret. To do this, we need the following facts, whose proofs are in Appendix A.9.

Proposition 2. *Following the Zooming Rule (27), we have*
1. *Each parent cube of measure μ is played at most $\frac{2(\Psi+D_{\mathbb{E}})^2\log(2T^2/\epsilon)}{\alpha^2[\log(\mu/\eta)]^2}$ episodes.*
2. *Under event $\widetilde{\mathcal{E}}_T$, each parent cube Q_t selected at episode t is a subset of $\mathcal{X}_{\delta}\left((1+2\alpha)\log\left(M_d\mu(Q_t)/\eta\right)\right)$.*

For cleaner writing, we set $\eta = 2^{-dI}$ for some positive integer I, and assume the event $\widetilde{\mathcal{E}}_T$ holds. By Proposition 2, we can regroup the regret in a similar way to that of Kleinberg et al. (2008). Let $\mathcal{K}_i$ be the collection of selected parent cubes such that for any $Q \in \mathcal{K}_i$, $\mu(Q) = 2^{-di}$ (dyadic cubes are always of these sizes). The sets $\mathcal{K}_i$ regroup the selected parent cubes by their size. By Proposition 2 (item 2), we know each parent cube in $\mathcal{K}_i$ is a subset

of $\mathcal{X}_\delta\left((1+2\alpha)\log\left(M_d 2^{-di}/\eta\right)\right)$. Since cubes in $\mathcal{K}_i$ are subsets of $\mathcal{X}_\delta\left((1+2\alpha)\log\left(M_d 2^{-di}/\eta\right)\right)$ and cubes in $\mathcal{K}_i$ are of measure 2^{-di}, we have

$$|\mathcal{K}_i| \le N_\delta\left((1+2\alpha)\log\left(M_d 2^{-di}/\eta\right), 2^{-i}\right), \tag{36}$$

where $|\mathcal{K}_i|$ is the number of cubes in $\mathcal{K}_i$. For a cube Q, let S_Q be the episodes where Q is played. With probability at least $1-2\epsilon$, we can regroup the regret as

$$\sum_{t=1}^{T}\sum_{i=1}^{M_d}\Delta_{t,i}^{\delta} \le \sum_{t=1}^{T}(1+2\alpha)M_d\log\left(M_d\mu(Q_t)/\eta\right) \tag{37}$$

$$\le \sum_{i=0}^{I-1}\sum_{Q\in\mathcal{K}_i}\sum_{t\in S_Q}(1+2\alpha)M_d\log\left(M_d 2^{-di}/\eta\right), \tag{38}$$

where (37) uses (33), (38) regroups the sum as argued above. Using Proposition 2, we can bound (38) by:

$$\sum_{i=0}^{I-1}\sum_{Q\in\mathcal{K}_i}\sum_{t\in S_Q}(1+2\alpha)M_d\log\left(M_d 2^{-di}/\eta\right)$$

$$\le \sum_{i=0}^{I-1}\sum_{Q\in\mathcal{K}_i}|S_Q|(1+2\alpha)M_d\log\left(\frac{M_d 2^{-di}}{\eta}\right)$$

$$\overset{①}{\le} \sum_{i=0}^{I-1}\sum_{Q\in\mathcal{K}_i}\frac{2(\Psi+D_{\mathcal{E}})^2\log(2T^2/\epsilon)}{\alpha^2\left[\log\left(2^{-di}/\eta\right)\right]^2} \tag{39}$$

$$\cdot(1+2\alpha)M_d\log\left(\frac{M_d 2^{-di}}{\eta}\right)$$

$$\le \sum_{i=0}^{I-1}N_\delta\left((1+2\alpha)\log\left(M_d 2^{-di}/\eta\right), 2^{-di}\right)$$

$$\cdot\frac{2(\Psi+D_{\mathcal{E}})^2\log(2T^2/\epsilon)}{\alpha^2\left[\log\left(2^{-di}/\eta\right)\right]^2}\cdot(1+2\alpha)M_d\log\left(\frac{M_d 2^{-di}}{\eta}\right) \tag{40}$$

$$\le \frac{2(1+2\alpha)M_d(\Psi+D_{\mathcal{E}})^2}{\alpha^2}\widetilde{N}_{\delta,\eta,\alpha}$$

$$\cdot\log(2T^2/\epsilon)\sum_{i=0}^{I-1}\frac{\log(M_d 2^{-di}/\eta)}{\left[\log(2^{-di}/\eta)\right]^2},$$

where ① uses item 1 in Proposition 2, (40) uses (36). Recall $\eta = 2^{-dI}$ for some positive integer I. We can use the above to prove Theorem 4, by using $\eta = 2^{-dI}$ and

$$\sum_{i=0}^{I-1}\frac{\log(M_d 2^{-di}/\eta)}{\left[\log(2^{-di}/\eta)\right]^2} = \sum_{i=0}^{I-1}\frac{\log M_d}{\left[\log\frac{2^{-di}}{\eta}\right]^2} + \sum_{i=0}^{I-1}\frac{1}{\log\frac{2^{-di}}{\eta}}$$

$$= \sum_{i=0}^{I-1}\frac{\log M_d}{d^2(\log 2)^2(I-i)^2} + \sum_{i=0}^{I-1}\frac{1}{d(\log 2)(I-i)} \tag{41}$$

$$= \mathcal{O}(1) + \mathcal{O}(\log I),$$

$$= \mathcal{O}(\log\log(1/\eta)),$$

where the first term in (41) is $\mathcal{O}(1)$ since $\sum_{i=1}^{\infty}\frac{1}{i^2} = \mathcal{O}(1)$ and the second term in (41) is $\mathcal{O}(\log I)$ by the order of a harmonic sum. The above analysis gives Theorem 4.

Theorem 4. *Choose positive integer I, and let $\eta = 2^{-Id}$. For $\epsilon > 0$ and $t \le T$, with probability $\ge 1-2\epsilon$, for any $\delta > |\mathcal{Q}_t|\eta$ such that δ is f-admissible, Algorithm 2 (with zooming rate α) admits t-episode δ-regret of:*

$$\mathcal{O}\left(\frac{1+2\alpha}{\alpha^2}M_d\Psi^2\widetilde{N}_{\delta,\eta,\alpha}\log\left(\frac{T}{\epsilon}\right)\log\log(1/\eta)\right), \tag{42}$$

where $\Psi = \mathcal{O}(\log(T/\epsilon)+\log(1/\eta))$, $\widetilde{N}_{\delta,\eta,\alpha}$ is defined in (35), and $\mathcal{O}$ omits constants. Since each episode plays M_d arms, the average δ-regret each arm incurs is independent of M_d.

When proving Theorem 4, the definition of $\widetilde{N}_{\delta,\eta,\alpha}$ is used in (40). For a more refined bound, we can instead use

$$\widetilde{N}'_{\delta,\eta,\alpha} := \sup_{\lambda\in(l_{\min},1]}N_\delta\left((1+2\alpha)\log\left(M_d\lambda^d/\eta\right),\lambda\right),$$

where $l_{\min}$ is the minimal possible cube edge length during the algorithm run. This replacement will not affect the argument. Some details and an example regarding this refinement are in Appendix A.10.

In Remark 3, we give an example of regret rate on $f(x) = 2\log\frac{1}{x}$, $x\in(0,1]$ with specific input parameters.

Remark 3. *Consider the (unbounded, BMO) function $f(x) = 2\log\frac{1}{x}$, $x\in(0,1]$. Pick $T \ge 20$. For some $t \le T$, the t-step δ-regret of Algorithm 2 is $\mathcal{O}(\text{poly-log}(t))$ while allowing $\delta = \mathcal{O}(1/T)$ and $\eta = \Theta\left(1/T^4\right)$. Intuitively, Algorithm 2 gets close to f^δ even if f^δ is very large. Details of this example can be found in Appendix A.10.*

7. Experiments

We deploy Algorithms 1 and 2 on the Himmelblau's function and the Styblinski-Tang function (arm space normalized to $[0,1)^2$, function range rescaled to $[0,10]$). The results are in Figure 3. We measure performance using traditional regret and δ-regret. Traditional regret can be measured because both functions are continuous, in addition to being BMO.

8. Discussion on Future Directions

8.1. Lower Bound

A classic trick to derive minimax lower bounds for (stochastic) bandit problems is the "needle-in-a-haystack." In this argument (Auer, 2002), we construct a hard problem instance, where one arm is only slightly better than the rest of the arms, making it hard to distinguish the best arm from the rest of the arms. This argument is also used in metric spaces (e.g., Kleinberg et al., 2008; Lu et al., 2019). This argument,

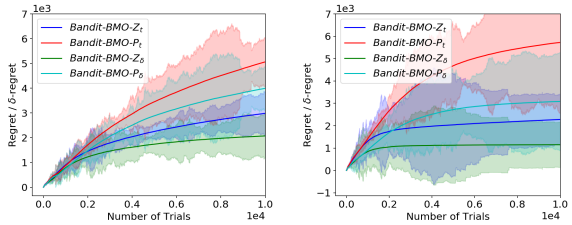

Figure 3. Algorithms 1 and 2 on Himmelblau's function (left) and Styblinski–Tang function (right). Each line is averaged over 10 runs. The shaded area represents one variance above and below the average regret. For the *Bandit-BMO-Z* algorithm, all arms played incur regret, and each episode has 4 arm trials in it. In the figures, *Bandit-BMO-*Z_δ (resp. *Bandit-BMO-*P_δ) plots the δ-regret ($\delta = 0.01$) for *Bandit-BMO-Z* (resp. *Bandit-BMO-P*). *Bandit-BMO-*Z_t (resp. *Bandit-BMO-*P_t) plots the traditional regret for *Bandit-BMO-Z* (resp. *Bandit-BMO-P*). For *Bandit-BMO-P* algorithm, we use $\epsilon = 0.01$, $\eta = 0.001$, total number of trials $T = 10000$. For *Bandit-BMO-Z* algorithm, we use $\alpha = 1$, $\epsilon = 0.01$, $\eta = 0.001$, number of episodes $T = 2500$, with four arm trials in each episode. Note that we have plotted trials (arm pulls) rather than episodes. The landscape of the test functions are in Appendix D.

however, is forbidden by the definition of δ-regret, since here, the set of good arms can have small measure, and will be ignored by definition. Hence, we need new insights to derive minimax lower bounds of bandit problems measured by δ-regret.

8.2. Singularities of Analytical Forms

In this paper, we investigate the bandit problem where the reward can have singularities in the arm space. A natural problem along this line is when the reward function has specific forms of singularities. For example, when the average reward can be written as $f(x) = \sum_{i=1}^{k} \frac{1}{(x-s_i)^{\alpha_i}}$ where s_i are the singularities and α_i are the "degree" of singularities. To continue leveraging the advantages of BMO function and John-Nirenberg inequalities, one might consider switching away from the Lebesgue measure and use decomposition results from classical analysis (e.g., Rochberg and Semmes, 1986).

9. Conclusion

We study the bandit problem when the (expected) reward is a BMO function. We develop tools for BMO bandits, and provide an algorithm that achieves poly-log δ-regret with high probability. Our result suggests that BMO functions can be optimized (with respect to δ-regret) even though they can be discontinuous and unbounded.

Acknowledgement

The authors thank Weicheng Ye and Jingwei Zhang for insightful discussions. The authors thank anonymous reviewers for valuable feedback.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agrawal, R. (1995). The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951.

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

Alon, N., Matias, Y., and Szegedy, M. (1999). The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147.

Assouad, P. (1983). Plongements Lipschitziens dans $\mathbb{R}^n$. *Bulletin de la Société Mathématique de France*, 111:429–448.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P. and Ortner, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

Auer, P., Ortner, R., and Szepesvári, C. (2007). Improved rates for the stochastic continuum-armed bandit problem. In *Conference on Computational Learning Theory*, pages 454–468. Springer.

Bickel, P. J. et al. (1965). On some robust estimates of location. *The Annals of Mathematical Statistics*, 36(3):847–858.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

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

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1.

Bubeck, S., Stoltz, G., and Yu, J. Y. (2011b). Lipschitz bandits without the Lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer.

Cope, E. W. (2009). Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control*, 54(6):1243–1253.

Cwikel, M., Sagher, Y., and Shvartsman, P. (2012). A new look at the John–Nirenberg and John–Strömberg theorems for BMO. *Journal of Functional Analysis*, 263(1):129–166.

Fefferman, R. (1979). Bounded mean oscillation on the polydisk. *Annals of Mathematics*, 110(3):395–406.

Garivier, A. and Cappé, O. (2011). The KL–UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, pages 359–376.

John, F. (1961). Rotation and strain. *Communications on Pure and Applied Mathematics*, 14(3):391–413.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *ACM Symposium on Theory of Computing*, pages 681–690. ACM.

Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704.

Krause, A. and Ong, C. S. (2011). Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455.

Krishnamurthy, A., Langford, J., Slivkins, A., and Zhang, C. (2019). Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *Conference on Learning Theory*, pages 2025–2027. PMLR.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lerner, A. K. (2013). The John–Nirenberg inequality with sharp constants. *Comptes Rendus Mathematique*, 351(11-12):463–466.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, pages 661–670. ACM.

Lu, S., Wang, G., Hu, Y., and Zhang, L. (2019). Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163.

Maillard, O.-A., Munos, R., and Stoltz, G. (2011). A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Conference On Learning Theory*, pages 497–514.

Martell, J. An easy proof of the John-Nirenberg inequality – math blog of Hyunwoo Will Kwon. http://willkwon.dothome.co.kr/index.php/archives/618, last accessed on 20/06/2020.

Medina, A. M. and Yang, S. (2016). No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Rochberg, R. and Semmes, S. (1986). A decomposition theorem for BMO and applications. *Journal of functional analysis*, 67(2):228–263.

Seldin, Y. and Slivkins, A. (2014). One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295.

Shamir, O. (2011). A variant of Azuma's inequality for martingales with sub-Gaussian tails. *arXiv preprint arXiv:1110.2392*.

Shao, H., Yu, X., King, I., and Lyu, M. R. (2018). Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, pages 8420–8429.

Slavin, L. and Vasyunin, V. (2017). The John–Nirenberg constant of BMO^p, $1 \leq p \leq 2$. *St. Petersburg Mathematical Journal*, 28(2):181–196.

Slivkins, A. (2014). Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*.

Stein, E. M. and Murphy, T. S. (1993). *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, volume 3. Princeton University Press.

Tao, T. and Vu, V. (2015). Random matrices: universality of local spectral statistics of non-hermitian matrices. *The Annals of Probability*, 43(2):782–874.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Vu, V. H. (2002). Concentration of non-Lipschitz functions and applications. *Random Structures & Algorithms*, 20(3):262–316.

Wang, T., Ye, W., Geng, D., and Rudin, C. (2019). Towards practical Lipschitz stochastic bandits. *arXiv preprint arXiv:1901.09277*.

Wanigasekara, N. and Yu, C. (2019). Nonparametric contextual bandits in an unknown metric space. In *Advances in Neural Information Processing Systems*.