Lipschitz Bandits with Batched Feedback

Yasong Feng
Shanghai Center for Mathematical Sciences
Fudan University
ysfeng20@fudan.edu.cn

Zengfeng Huang*
School of Data Science
Fudan University
huangzf@fudan.edu.cn

Tianyu Wang*
Shanghai Center for Mathematical Sciences
Fudan University
wangtianyu@fudan.edu.cn

Abstract

In this paper, we study Lipschitz bandit problems with batched feedback, where the expected reward is Lipschitz and the reward observations are communicated to the player in batches. We introduce a novel landscape-aware algorithm, called Batched Lipschitz Narrowing (BLiN), that optimally solves this problem. Specifically, we show that for a T-step problem with Lipschitz reward of zooming dimension d_z, our algorithm achieves theoretically optimal (up to logarithmic factors) regret rate $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ using only $\mathcal{O}\left(\log\log T\right)$ batches. We also provide complexity analysis for this problem. Our theoretical lower bound implies that $\Omega(\log\log T)$ batches are necessary for any algorithm to achieve the optimal regret. Thus, BLiN achieves optimal regret rate using minimal communication.

1 Introduction

Multi-Armed Bandit (MAB) algorithms aim to exploit the good options while explore the decision space. These algorithms and methodologies find successful applications in artificial intelligence and reinforcement learning [e.g., 37]. While the classic MAB setting assumes that the rewards are immediately observed after each arm pull, real-world data often arrives in different patterns. For example, observations from clinical trials are often be collected in a batched fashion [34]. Another example is from online advertising, where strategies are tested on multiple customers at the same time [10]. In such cases, any observation-dependent decision-making should comply with this data-arriving pattern, including MAB algorithms.

In this paper, we study the Lipschitz bandit problem with batched feedback – a MAB problem where the expected reward is Lipschitz and the reward observations are communicated to the player in batches. In such settings, rewards are communicated only at the end of the batches, and the algorithm can only make decisions based on information up to the previous batch. Existing Lipschitz bandit algorithms heavily rely on timely access to the reward samples, since the partition of arm space may change at any time. Therefore, they can not solve the batched feedback setting. To address this difficulty, we present a novel adaptive algorithm for Lipschitz bandits with communication constraints, named *Batched Lipschitz Narrowing* (BLiN). BLiN learns the landscape of the reward by adaptively narrowing the arm set, so that regions of high reward are more frequently played. Also, BLiN determines the data collection procedure adaptively, so that only very few data communications are needed.

*Corresponding authors

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The above BLiN procedure achieves optimal regret rate $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ (d_z is the zooming dimension [26, 15]), and can be implemented in a clean and easy-to-implement form. In addition to achieving the optimal regret rate, BLiN is also optimal in the following senses:

- BLiN's communication complexity is optimal. BLiN only needs $\mathcal{O}(\log\log T)$ rounds of communications to achieve the optimal regret rate (Theorem 2), and no algorithm can achieve this rate with fewer than $\Omega(\log\log T)$ rounds of communications (Corollary 2).
- BLiN's time complexity is optimal (Remark 1): if the arithmetic operations and sampling are of complexity $\mathcal{O}(1)$, then the time complexity of BLiN is $\mathcal{O}(T)$, which improve the best known time complexity $\mathcal{O}(T\log T)$ for Lipschitz bandit problems [15].

1.1 Settings & Preliminaries

For a Lipschitz bandit problem (with communication constraints), the arm set is a compact doubling metric space $(\mathcal{A}, d_{\mathcal{A}})$. The expected reward $\mu : \mathcal{A} \to \mathbb{R}$ is 1-Lipschitz with respect to the metric $d_{\mathcal{A}}$, that is, $|\mu(x_1) - \mu(x_2)| \le d_{\mathcal{A}}(x_1, x_2)$ for any $x_1, x_2 \in \mathcal{A}$.

At time $t \le T$, the learning agent pulls an arm $x_t \in \mathcal{A}$ that yields a reward sample $y_t = \mu(x_t) + \epsilon_t$, where ϵ_t is a mean-zero independent sub-Gaussian noise. Without loss of generality, we assume that $\epsilon_t \sim \mathcal{N}(0,1)$, since generalizations to other sub-Gaussian noises are not hard.

Similar to most bandit learning problems, the agent seeks to minimize regret in the batched feedback environment. The regret is defined as $R(T) = \sum_{t=1}^{T}\left(\mu^* - \mu(x_t)\right)$, where μ^* denotes $\max_{x\in\mathcal{A}}\mu(x)$. For simplicity, we define $\Delta_x = \mu^* - \mu(x)$ (called optimality gap of x) for all $x \in \mathcal{A}$.

1.1.1 Doubling Metric Spaces and the $([0,1]^d, \|\cdot\|_\infty)$ Metric Space

By the Assouad's embedding theorem [6], the (compact) doubling metric space $(\mathcal{A}, d_{\mathcal{A}})$ can be embedded into a Euclidean space with some distortion of the metric; See [43] for more discussions in a machine learning context. Due to existence of such embedding, the metric space $([0,1]^d, \|\cdot\|_\infty)$, where metric balls are hypercubes, is sufficient for the purpose of our paper. For the rest of the paper, we will use hypercubes in algorithm design for simplicity, while our algorithmic idea generalizes to other doubling metric spaces.

1.1.2 Zooming Numbers and Zooming Dimensions

An important concept for bandit problems in metric spaces is the zooming number and the zooming dimension [26, 14, 39], which we discuss now.

Define the set of r-optimal arms as $S(r) = \{x \in \mathcal{A} : \Delta_x \le r\}$. For any $r = 2^{-i}$, the decision space $[0,1]^d$ can be equally divided into 2^{di} cubes with edge length r, which we call *standard cubes* (also referred to as dyadic cubes). The r-zooming number is defined as

$$N_r := \#\{C : C \text{ is a standard cube with edge length } r \text{ and } C \subset S(16r)\}.$$

In words, N_r is the r-packing number of the set $S(16r)$ in terms of standard cubes. The zooming dimension is then defined as

$$d_z := \min\{d \ge 0 : \exists a > 0,\ N_r \le ar^{-d},\ \forall r = 2^{-i} \text{ for some } i \in \mathbb{N}\}.$$

Moreover, we define the zooming constant C_z as the minimal a to make the above inequality true for d_z, $C_z = \min\{a > 0 :\ N_r \le ar^{-d_z},\ \forall r = 2^{-i} \text{ for some } i \in \mathbb{N}\}$.

Zooming dimension d_z can be significantly smaller than ambient dimension d and can be zero. For a simple example, consider a problem with ambient dimension $d = 1$ and expected reward function $\mu(x) = x$ for $0 \le x \le 1$. Then for any $r = 2^{-i}$ with $i \ge 4$, we have $S(16r) = [1 - 16r, 1]$ and $N_r = 16$. Therefore, for this problem the zooming dimension equals to 0, with zooming constant $C_z = 16$.

1.2 Batched feedback pattern and our results

In the batched feedback setting, for a T-step game, the player determines a grid $\mathcal{T} = \{t_0, \cdots, t_B\}$ *adaptively*, where $0 = t_0 < t_1 < \cdots < t_B = T$ and $B \ll T$. During the game, reward observations are communicated to the player only at the grid points $t_1, \cdots, t_B$. As a consequence, for any time t in the j-th batch, that is, $t_{j-1} < t \leq t_j$, the reward y_t cannot be observed until time t_j, and the decision made at time t depends only on rewards up to time t_{j-1}. The determination of the grid $\mathcal{T}$ is adaptive in the sense that the player chooses each grid point $t_j \in \mathcal{T}$ based on the operations and observations up to the previous point t_{j-1}.

In this work, we present BLiN algorithm to solve Lipschitz bandits under batched feedback. During the learning procedure, BLiN detects and eliminates the 'bad area' of the arm set in batches and partition the remaining area according to an approporiate *edge-length sequence*. Our first theoretical upper bound is that with simple Doubling Edge-length Sequence $r_m = 2^{-m+1}$, BLiN achieves optimal regret rate $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ by using $\mathcal{O}(\log T)$ batches.

Theorem 1. *With probability exceeding $1 - \frac{2}{T^6}$, the T-step total regret $R(T)$ of BLiN with Doubling Edge-length Sequence (D-BLiN) satisfies*

$$R(T) \lesssim T^{\frac{d_z+1}{d_z+2}} \cdot (\log T)^{\frac{1}{d_z+2}},$$

where d_z is the zooming dimension of the problem instance. In addition, D-BLiN only needs no more than $\mathcal{O}(\log T)$ rounds of communications to achieve this regret rate. Here and henceforth, $\lesssim$ only omits constants.

While D-BLiN is efficient for batched Lipschitz bandits, its communication complexity is not optimal. We then propose a new edge-length sequence, which we call Appropriate Combined Edge-length Sequence (ACE Sequence) to improve the algorithm. The idea behind this sequence is that by appropriately combining some batches, the algorithm can achieve better communication bound without incurring increased regret. As we shall see, BLiN with ACE Sequence (A-BLiN) achieves regret rate $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ with only $\mathcal{O}(\log\log T)$ batches.

Theorem 2. *With probability exceeding $1 - \frac{2}{T^6}$, the T-step total regret $R(T)$ of A-BLiN satisfies*

$$R(T) \lesssim T^{\frac{d_z+1}{d_z+2}} \cdot (\log T)^{\frac{1}{d_z+2}} \cdot \log\log T,$$

where d_z is the zooming dimension of the problem instance. In addition, Algorithm 1 only needs no more than $\mathcal{O}(\log\log T)$ rounds of communications to achieve this regret rate.

As a comparison, seminal works [26, 39, 15] show that the optimal regret bound for Lipschitz bandits without communications constraints, where the reward observations are immediately observable after each arm pull, is $R(T) \lesssim T^{\frac{d_z+1}{d_z+2}} \cdot (\log T)^{\frac{1}{d_z+2}}$ in terms of zooming dimension, and

$$R(T) \lesssim R_z(T) \triangleq \inf_{r_0} \left\{ r_0 T + \sum_{r=2^{-i}, r \geq r_0} \frac{N_r}{r} \log T \right\} \tag{1}$$

in terms of zooming number. Therefore, A-BLiN achieves optimal regret rate of Lipschitz bandits by using very few batches.

Moreover, our lower-bound analysis shows that $\Omega(\log\log T)$ batches are necessary for any algorithm to achieve the optimal regret rate. Thus, BLiN is optimal in terms of both regret and communication.

1.3 Related Works

The history of the Multi-Armed Bandit (MAB) problem can date back to Thompson [42]. Solvers for this problems include the UCB algorithms [28, 4, 7], the arm elimination method [20, 32, 36], the ϵ-greedy strategy [7, 40], the exponential weights and mirror descent framework [8].

Recently, with the prevalence of distributed computing and large-scale field experiments, the setting of batched feedback has captured attention [e.g., 17]. Perchet et al. [33] mainly consider batched bandit with two arms, and a matching lower bound for static grid is proved. It was then generalized by Gao et al. [21] to finite-armed bandit problems. In their work, the authors designed an elimination method for finite-armed bandit problem and proved matching lower bounds for both static and adaptive grid. Soon afterwards, Zhang et al. [46] studies inference for batched bandits. Esfandiari et al. [19] studies batched linear bandits and batched adversarial bandits. Han et al. [22] and Ruan et al. [35] provide solutions for batched contextual linear bandits. Li and Scarlett [29] studies batched Gaussian process bandits. Batched dueling bandits have also been studied by [2]. Parallel to the regret control regime, best arm identification with limited number of batches was studied in [1] and [23]. Top-k arm identification in the collaborative learning framework is also closely related to the batched setting, where the goal is to minimize the number of iterations (or communication steps) between agents. In this setting, tight bounds have been obtained in the recent works [41, 24]. Yet the problem of Lipschitz bandit with communication constraints remains unsolved.

The Lipschitz bandit problem is important in its own stand. The Lipschitz bandit problem was introduced as "continuum-armed bandits" [3], where the arm space is a compact interval. Along this line, bandits that are Lipschitz (or Hölder) continuous have been studied. For this problem, Kleinberg [25] proves a $\Omega(T^{2/3})$ lower bound and introduced a matching algorithm. Under extra conditions on top of Lipschitzness, regret rate of $\widetilde{\mathcal{O}}(T^{1/2})$ was achieved [9, 18]. For general (doubling) metric spaces, the Zooming bandit algorithm [26] and the Hierarchical Optimistic Optimization (HOO) algorithm [15] were developed. In more recent years, some attention has been focused on Lipschitz bandit problems with certain extra structures. To name a few, Bubeck et al. [16] study Lipschitz bandits for differentiable rewards, which enables algorithms to run without explicitly knowing the Lipschitz constants. Wang et al. [44] studied discretization-based Lipschitz bandit algorithms from a Gaussian process perspective. Magureanu et al. [31] derive a new concentration inequality and study discrete Lipschitz bandits. The idea of robust mean estimators [11, 5, 13] was applied to the Lipschitz bandit problem to cope with heavy-tail rewards, leading to the development of a near-optimal algorithm for Lipschitz bandit with heavy-tailed rewards [30]. Lipschitz bandits where a clustering is used to infer the underlying metric, has been studied by [45]. Contextual Lipschitz bandits have also been studied by [39] and [27]. Yet all of the existing works for Lipschitz bandits assume that the reward sample is immediately observed after each arm pull, and none of them solve the Lipschitz bandit problem with communication constraints.

This paper is organized as follows. In section 2, we introduce the BLiN algorithm and give a visual illustration of the algorithm procedure. In section 3, we prove that BLiN with ACE Sequence achieves the optimal regret rate using only $\mathcal{O}(\log\log T)$ rounds of communications. Section 4 provides information-theoretical lower bounds for Lipschitz bandits with communication constraints, which shows that BLiN is optimal in terms of both regret and rounds of communications. Experimental results are presented in Section 5.

2 Algorithm

With communication constraints, the agent's knowledge about the environment does not accumulate within each batch. This characteristic of the problem suggests a 'uniform' type algorithm – we shall treat each step within the same batch equally. Following this intuition, in each batch, we uniformly play the remaining arms, and then eliminate arms of low reward after the observations are communicated. Next we describe the uniform play rule and the arm elimination rule.

Uniform Play Rule: At the beginning of each batch m, a collection of subsets of the arm space $\mathcal{A}_m = \{C_{m,1}, C_{m,2}, \cdots, C_{m,|\mathcal{A}_m|}\}$ is constructed. This collection of subset $\mathcal{A}_m$ consists of standard cubes, and all cubes in $\mathcal{A}_m$ have the same edge length r_m. We will detail the construction of $\mathcal{A}_m$ when we describe the arm elimination rule. We refer to cubes in $\mathcal{A}_m$ as active cubes of batch m.

During batch m, each cube in $\mathcal{A}_m$ is played $n_m := \frac{16\log T}{r_m^2}$ times, where T is the total time horizon. More specifically, within each $C \in \mathcal{A}_m$, arms $x_{C,1}, x_{C,2}, \cdots, x_{C,n_m} \in C$ are played.[2] The reward samples $\{y_{C,1}, y_{C,2}, \cdots, y_{C,n_m}\}_{C\in\mathcal{A}_m}$ corresponding to $\{x_{C,1}, x_{C,2}, \cdots, x_{C,n_m}\}_{C\in\mathcal{A}_m}$ will be collected at the end of the this batch.

Arm Elimination Rule: At the end of batch m, information from the arm pulls is collected, and we estimate the reward of each $C \in \mathcal{A}_m$ by $\widehat{\mu}_m(C) = \frac{1}{n_m}\sum_{i=1}^{n_m} y_{C,i}$. Cubes of low estimated rewards are then eliminated, according to the following rule: a cube $C \in \mathcal{A}_m$ is eliminated if $\widehat{\mu}_m^{\max} - \widehat{\mu}_m(C) \geq 4r_m$, where $\widehat{\mu}_m^{\max} := \max_{C\in\mathcal{A}_m}\widehat{\mu}_m(C)$. After necessary removal of "bad cubes", each cube in $\mathcal{A}_m$ that survives the elimination is equally partitioned into $\left(\frac{r_m}{r_{m+1}}\right)^d$ subcubes of edge

[2]One can arbitrarily play $x_{C,1}, x_{C,2}, \cdots, x_{C,n_m}$ as long as $x_{C,i} \in C$ for all i.

length r_{m+1}, where r_{m+1} is predetermined. These cubes (of edge length r_{m+1}) are collected to construct $\mathcal{A}_{m+1}$, and the learning process moves on to the next batch. Appropriate rounding may be required to ensure the ratio $\frac{r_m}{r_{m+1}}$ is an integer. See Remark 2 for more details.

The learning process is summarized in Algorithm 1.

Algorithm 1 Batched Lipschitz Narrowing (BLiN)

Input. Arm set $\mathcal{A} = [0,1]^d$; time horizon T.
Initialization Number of batches B; Edge-length sequence $\{r_m\}_{m=1}^{B+1}$; The first grid point $t_0 = 0$; Equally partition $\mathcal{A}$ to r_1^d subcubes and define $\mathcal{A}_1$ as the collection of these subcubes.
Compute $n_m = \frac{16 \log T}{r_m^2}$ for $m = 1, \cdots, B+1$.
for $m = 1, 2, \cdots, B$ **do**
  For each cube $C \in \mathcal{A}_m$, play arms $x_{C,1}, \cdots x_{C,n_m}$ from C.
  Collect the rewards of all pulls up to t_m. Compute the average payoff $\widehat{\mu}_m(C) = \frac{\sum_{i=1}^{n_m} y_{C,i}}{n_m}$ for each cube $C \in \mathcal{A}_m$. Find $\widehat{\mu}_m^{max} = \max_{C \in \mathcal{A}_m} \widehat{\mu}(C)$.
  For each cube $C \in \mathcal{A}_m$, eliminate C if $\widehat{\mu}_m^{max} - \widehat{\mu}_m(C) > 4r_m$. Let $\mathcal{A}_m^+$ be set of cubes not eliminated.
  Compute $t_{m+1} = t_m + (r_m/r_{m+1})^d \cdot |\mathcal{A}_m^+| \cdot n_{m+1}$. If $t_{m+1} \geq T$ or $m = B$ then **break**.
  Equally partition each cube in $\mathcal{A}_m^+$ into $(r_m/r_{m+1})^d$ subcubes and define $\mathcal{A}_{m+1}$ as the collection of these subcubes. /**See Remark 2 for more details on cases where $(r_m/r_{m+1})^d$ is not an integer.**/
end for
Cleanup: Arbitrarily play the remaining arms until all T steps are used.

The following theorem gives regret and communication upper bound of BLiN with Doubling Edge-length Sequence $r_m = 2^{-m+1}$ (see Appendix B for proof). Note that this result implies Theorem 1.

Theorem 3. *With probability exceeding $1 - \frac{2}{T^6}$, the T-step total regret $R(T)$ of BLiN with Doubling Edge-length Sequence (D-BLiN) satisfies*

$$R(T) \leq 528 (\log T)^{\frac{1}{d_z+2}} \cdot T^{\frac{d_z+1}{d_z+2}},$$

where d_z is the zooming dimension of the problem instance. In addition, D-BLiN only needs no more than $\frac{\log T - \log\log T}{d_z+2} + 2$ rounds of communications to achieve this regret rate.

Although D-BLiN efficiently solves batched Lipschitz bandits, its simple partition strategy leads to suboptimal communication complexity. Now we show that by approporiately combining some batches, BLiN achieves the optimal communication bound, without incurring increasing regret. Specifically, we introduce the following edge-length sequence, which we call ACE Sequence.

Definition 1. *For a problem with ambient dimension d, zooming dimension d_z and time horizon T, we denote $c_1 = \frac{d_z+1}{(d+2)(d_z+2)} \log \frac{T}{\log T}$ and $c_{i+1} = \eta c_i$ for any $i \geq 1$, where $\eta = \frac{d+1-d_z}{d+2}$. Then the Appropriate Combined Edge-length Sequence $\{r_m\}$ is defined by $r_m = 2^{-\sum_{i=1}^m c_i}$ for any $m \geq 1$.*

We show that BLiN with ACE Sequence (A-BLiN) obtains an improved communication complexity, thus proves Theorem 2.

Theorem 4. *With probability exceeding $1 - \frac{2}{T^6}$, the T-step total regret $R(T)$ of Algorithm 1 satisfies*

$$R(T) \leq \left(\frac{128 C_z}{\log \frac{d+2}{d+1-d_z}} \cdot \log\log T + 8e \right) \cdot T^{\frac{d_z+1}{d_z+2}} (\log T)^{\frac{1}{d_z+2}}, \tag{2}$$

where d_z is the zooming dimension of the problem instance. In addition, Algorithm 1 only needs no more than $\frac{\log\log T}{\log \frac{d+2}{d+1-d_z}} + 1$ rounds of communications to achieve this regret rate.

The partition and elimination process of a real A-BLiN run is in Figure 1. In the i-th subgraph, the white cubes are those remaining after the $(i-1)$-th batch. In this experiment, we set $\mathcal{A} = [0,1]^2$, and the optimal arm is $x^* = (0.8, 0.7)$. Note that x^* is not eliminated during the game. More details of this experiment are in Section 5.

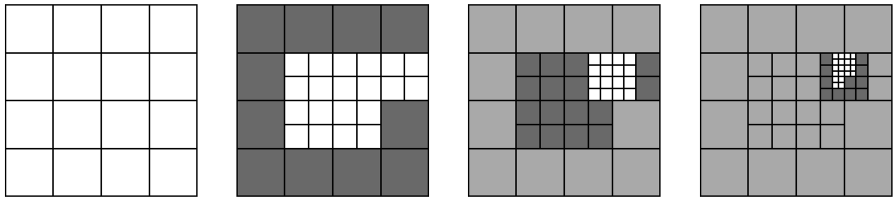


Figure 1: Partition and elimination process of A-BLiN. The i-th subfigure shows the pattern before the i-th batch, which is from a real A-BLiN run on the reward function defined in Section 5. Dark gray cubes are those eliminated in the most recent batch, while the light gray ones are those eliminated in earlier batches. For the total time horizon $T = 80000$, A-BLiN needs 4 rounds of communications. For this experiment, $r_1 = \frac{1}{4},\ r_2 = \frac{1}{8},\ r_3 = \frac{1}{16},\ r_4 = \frac{1}{32},$ which is the ACE sequence (rounded as in Remark 2) for $d = 2$ and $d_z = 0$.

Remark 1 (Time and space complexity)**.** *The time complexity of our algorithm is $\mathcal{O}(T)$, which is better than the state of the art $\mathcal{O}(T \log T)$ in [15]. This is because that the running time of a batch j is of order $\mathcal{O}(l_j)$, where $l_j = t_j - t_{j-1}$ is number of samples in batch j. Since $\sum_j l_j = T$, the time complexity of BLiN is $\mathcal{O}(T)$. Besides, the space complexity of BLiN is also improved, because we do not need to store information of cubes in previous batches.*

3 Regret Analysis of A-BLiN

In this section, we provide regret analysis for A-BLiN. The highlight of the finding is that $\mathcal{O}(\log\log T)$ batches are sufficient to achieve optimal regret rate of $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}}\right)$, as summarized in Theorem 4.

To prove Theorem 4, we first show that the estimator $\widehat{\mu}$ is concentrated to the true expected reward μ (Lemma 1), and the optimal arm survives all eliminations with high probability (Lemma 2). In the following analysis, we let B_{stop} be the total number of batches of the BLiN run.

Lemma 1. *Define*

$$\mathcal{E} := \left\{ |\mu(x) - \widehat{\mu}_m(C)| \le r_m + \sqrt{\frac{16\log T}{n_m}}, \forall 1 \le m \le B_{\text{stop}} - 1,\ \forall C \in \mathcal{A}_m,\ \forall x \in C \right\}.$$

It holds that $\mathbb{P}(\mathcal{E}) \ge 1 - 2T^{-6}$.

Lemma 2. *Under event $\mathcal{E}$, the optimal arm $x^* = \arg\max \mu(x)$ is not eliminated after the first $B_{\text{stop}} - 1$ batches.*

Based on these results, we show the cubes that survive elimination are of high reward.

Lemma 3. *Under event $\mathcal{E}$, for any $1 \le m \le B_{\text{stop}}$, any $C \in \mathcal{A}_m$ and any $x \in C$, Δ_x satisfies*

$$\Delta_x \le 8r_{m-1}. \tag{3}$$

The proofs of Lemma 1-3 are in Appendix A. We are now ready to prove the theorem.

Proof of Theorem 4. Let R_m denote regret of the m-th batch. Fixing any positive number B, the total regret $R(T)$ can be divided into two parts: $R(T) = \sum_{m \le B} R_m + \sum_{m > B} R_m$. In the following, we bound these two parts separately and then determine B to obtain the upper bound of the total regret. Moreover, we show A-BLiN uses only $\mathcal{O}(\log\log T)$ rounds of communications to achieve the optimal regret.

Recall that $\mathcal{A}_m$ is set of the active cubes in the m-th batch. According to Lemma 3, for any $x \in \cup_{C \in \mathcal{A}_m} C$, we have $\Delta_x \le 8r_{m-1}$. Let $\mathcal{A}_m^+$ be set of cubes not eliminated in batch m. Then any cube in $\mathcal{A}_{m-1}^+$ is a subset of $S(8r_{m-1})$, and thus

$$|\mathcal{A}_{m-1}^+| \le N_{r_{m-1}} \le C_z r_{m-1}^{-d_z}. \tag{4}$$

By definition, $r_m = r_{m-1}2^{-c_m}$, so

$$|\mathcal{A}_m| = 2^{dc_m}|\mathcal{A}^+_{m-1}|. \tag{5}$$

The total regret of the m-th batch is

$$\begin{aligned} R_m = \sum_{C\in\mathcal{A}_m}\sum_{i=1}^{n_m}\Delta_{x_{C,i}} &\leq |\mathcal{A}_m|\cdot\frac{16\log T}{r_m^2}\cdot 8r_{m-1} \overset{(i)}{=} 2^{dc_m}|\mathcal{A}^+_{m-1}|\cdot\frac{16\log T}{r_m^2}\cdot 8r_{m-1} \\ &\overset{(ii)}{\leq} 2^{dc_m}\cdot C_z r_{m-1}^{-d_z+1}\cdot\frac{128\log T}{r_m^2} \overset{(iii)}{=} 2^{(\sum_{i=1}^{m-1}c_i)(d_z+1)+c_m(d+2)}\cdot 128C_z\log T, \end{aligned}$$

where (i) follows from (5), (ii) follows from (4), and (iii) follows from the definition of ACE Sequence. Define $C_m = (\sum_{i=1}^{m-1}c_i)(d_z+1)+c_m(d+2)$. Since $c_m = c_{m-1}\cdot\frac{d+1-d_z}{d+2}$, calculation shows that $C_m = (\sum_{i=1}^{m-2}c_i)(d_z+1)+c_{m-1}(d+2)+c_{m-1}(d_z+1-d-2)+c_m(d+2) = C_{m-1}$. Thus for any m, we have $C_m = C_1 = c_1(d+2)$. Hence,

$$R_m \leq 2^{c_1(d+2)}\cdot 128C_z\log T = T^{\frac{d_z+1}{d_z+2}}\cdot 128C_z(\log T)^{\frac{1}{d_z+2}}. \tag{6}$$

The inequality (6) holds even if the m-th batch does not exist (where we let $R_m = 0$) or is not completed. Thus we obtain the first upper bound $\sum_{m\leq B}R_m \leq T^{\frac{d_z+1}{d_z+2}}\cdot 128C_z\cdot B(\log T)^{\frac{1}{d_z+2}}$. Lemma 3 implies that any arm x played after the first B batches satisfies $\Delta_x\leq 8r_B$, so the total regret after B batches is bounded by

$$\begin{aligned} \sum_{m>B}R_m \leq& 8r_B\cdot T = 8T\cdot 2^{-\sum_{i=1}^{B}c_i} = 8T\cdot 2^{-c_1(\frac{1-\eta^B}{1-\eta})} \\ =& 8T^{\frac{d_z+1}{d_z+2}}(\log T)^{\frac{1}{d_z+2}}\cdot(T/\log T)^{\frac{\eta^B}{d_z+2}} \leq 8T^{\frac{d_z+1}{d_z+2}}(\log T)^{\frac{1}{d_z+2}}\cdot T^{\frac{\eta^B}{d_z+2}}. \end{aligned}$$

Therefore, the total regret $R(T)$ satisfies

$$R(T) = \sum_{m\leq B}R_m + \sum_{m>B}R_m \leq \left(128C_z\cdot B + 8T^{\frac{\eta^B}{d_z+2}}\right)\cdot T^{\frac{d_z+1}{d_z+2}}(\log T)^{\frac{1}{d_z+2}}.$$

This inequality holds for any positive B. Then by choosing $B^* = \frac{\log\log T-\log(d_z+2)}{\log\frac{1}{\eta}} = \frac{\log\log T-\log(d_z+2)}{\log\frac{d+2}{d+1-d_z}}$, we have $\frac{\eta^{B^*}}{d_z+2} = \frac{1}{\log T}$ and

$$R(T) \leq \left(\frac{128C_z\log\log T}{\log\frac{d+2}{d+1-d_z}} + 8e\right)\cdot T^{\frac{d_z+1}{d_z+2}}(\log T)^{\frac{1}{d_z+2}}.$$

The above analysis implies that we can achieve the optimal regret rate $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ by letting the *for-loop* run B^* times and finishing the remaining rounds in the *Cleanup* step. In other words, B^*+1 rounds of communications are sufficient for A-BLiN to achieve the regret bound (2). $\square$

Remark 2. *The quantity $\frac{r_m}{r_{m+1}}$ in line 9 of Algorithm 1 may not be integers for some m. Thus, in practice we denote $\alpha_n = \lfloor\sum_{i=1}^n c_i\rfloor$, $\beta_n = \lceil\sum_{i=1}^n c_i\rceil$, and define rounded ACE Sequence $\{\widetilde{r}_m\}_{m\in\mathbb{N}}$ by $\widetilde{r}_m = 2^{-\alpha_k}$ for $m = 2k-1$ and $\widetilde{r}_m = 2^{-\beta_k}$ for $m = 2k$. Then the total regret can be divided as $R(T) = \sum_{1\leq k\leq B^*}R_{2k-1} + \sum_{1\leq k\leq B^*}R_{2k} + \sum_{m>2B^*}R_m$. For the first part we have $\widetilde{r}_{2k-2}\leq r_{k-1}$ and $\widetilde{r}_{2k-1}\geq r_k$, while for the second part we have $\frac{\widetilde{r}_{2k-1}}{\widetilde{r}_{2k}} = 2$. Therefore, by similar argument to the proof of Theorem 4, we can bound these three parts separately, and conclude that BLiN with rounded ACE sequence achieves the optimal regret bound $\widetilde{\mathcal{O}}(T^{\frac{d_z+1}{d_z+2}})$ by using only $\mathcal{O}(\log\log T)$ rounds of communications. The details are in Appendix C.*

4 Lower Bounds

In this section, we present lower bounds for Lipschitz bandits with batched feedback, which in turn gives communication lower bounds for all Lipschitz bandit algorithms. Our lower bounds depend on the rounds of communications B. When B is sufficiently large, our results match the lower bound for the vanilla Lipschitz bandit problem $\widetilde{\Theta}(R_z(T))$ ($R_z(T)$ is defined in Eq. 1). More importantly, this dependency on B gives the minimal rounds of communications needed to achieve optimal regret bound for all Lipschitz bandit algorithms, which is summarized in Corollary 2. Since this lower bound matches the upper bound presented in Theorem 4, BLiN optimally solves Lipschitz bandits with minimal communication.

4.1 Proof Outline

Similar to most lower bound proofs, we need to construct problem instances that are difficult to differentiate. What's different is that we need to carefully integrate batched feedback pattern [33] with the Lipschitz payoff reward [39, 30]. To capture the adaptivity in grid determination, we construct "static reference communication grids" to remove the stochasticity in grid selection [1, 21]. Below, we first consider the static grid case, where the grid is predetermined. This static grid case will provide intuition for the adaptive and more general case.

The expected reward functions of these instances are constructed as follows: we choose some 'position' and 'height', such that the expected reward function obtains local maximum of the specified 'height' at the specified 'position'. We will use the word 'peak' to refer to the local maxima. The following theorem presents the lower bound for the static grid case.

Theorem 5 (Lower Bound for Static Grid)**.** *Consider Lipschitz bandit problems with time horizon T such that the grid of reward communication $\mathcal{T}$ is static and determined before the game. If B rounds of communications are allowed, then for any policy π, there exists a problem instance such that*

$$\mathbb{E}[R_T(\pi)] \geq c \cdot (\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} \cdot R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}},$$

where $c > 0$ is a numerical constant independent of B, T, π and $\mathcal{T}$, $R_z(T)$ is defined in (1), and d is the dimension of the arm space.

To prove Theorem 5, we first show that for any $k > 1$ there exists an instance such that $\mathbb{E}[R_T(\pi)] \geq \frac{t_k}{t_{k-1}^{\frac{1}{d+2}}}$. Fixing $k > 1$, we let $r_k = \frac{1}{t_{k-1}^{\frac{1}{d+2}}}$ and $M_k := t_{k-1} r_k^2 = \frac{1}{r_k^d}$. Then we construct a set of problem instances $\mathcal{I}_k = \{I_{k,1}, \cdots, I_{k,M_k}\}$, such that the gap between the highest peak and the second highest peak is about r_k for every instance in $\mathcal{I}_k$.

Based on this construction, we prove that no algorithm can distinguish instances in $\mathcal{I}_k$ from one another in the first $(k-1)$ batches, so the worst-case regret is at least $r_k t_k$, which gives the inequality we need. For the first batch $(0, t_1]$, we can easily construct a set of instances where the worst-case regret is at least t_1, since no information is available during this time. Thus, there exists a problem instance such that $\mathbb{E}[R_T(\pi)] \gtrsim \max\left\{t_1, \frac{t_2}{t_1^{\frac{1}{d+2}}}, \cdots, \frac{t_B}{t_{B-1}^{\frac{1}{d+2}}}\right\}$. Since $0 < t_1 < \cdots < t_B = T$, the inequality in Theorem 5 follows.

As a result of Theorem 5, we can derive the minimum rounds of communications needed to achieve optimal regret bound for Lipschitz bandit problem, which is stated in Corollary 1.

Corollary 1. *Any Lipschitz bandit algorithm needs $\Omega(\log \log T)$ rounds of communications to achieve the optimal regret rate, for the case that the times of reward communication are predetermined and static.*

The detailed proof of Theorem 5 and Corollary 1 are deferred to Appendix D and E.

4.2 Communication Lower Bound for BLiN

So far we have derived lower bounds for the static grid case. Yet there is a gap between the static and the adaptive case. We will close this gap in the following Theorem.

Theorem 6 (Lower Bound for Adaptive Grid)**.** *Consider Lipschitz bandit problems with time horizon T such that the grid of reward communication $\mathcal{T}$ is adaptively determined by the player. If B rounds of communications are allowed, then for any policy π, there exists an instance such that*

$$\mathbb{E}\left[R_T(\pi)\right] \geq c\frac{1}{B^2}(\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}},$$

where $c > 0$ is a numerical constant independent of B, T, π and $\mathcal{T}$, $R_z(T)$ is defined in (1), and d is the dimension of the arm space.

To prove Theorem 6, we consider a reference static grid $\mathcal{T}_r = \{T_0, T_1, \cdots, T_B\}$, where $T_j = T^{\frac{1-\varepsilon^j}{1-\varepsilon^B}}$ for $\varepsilon = \frac{1}{d+2}$. We construct a series of 'worlds', denoted by $\mathcal{I}_1, \cdots, \mathcal{I}_B$. Each world is a set of problem instances, and each problem instance in world $\mathcal{I}_j$ is defined by peak location set $\mathcal{U}_j$ and basic height r_j, where the sets $\mathcal{U}_j$ and quantities r_j for $1 \leq j \leq B$ are presented in Appendix F.

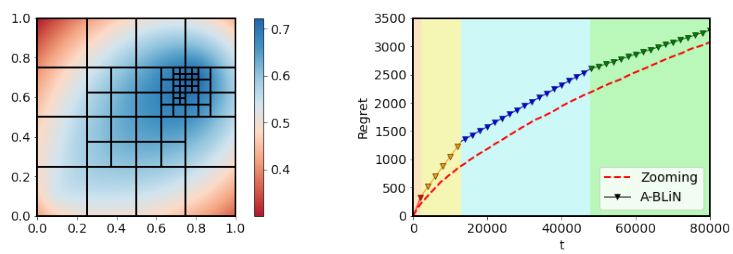
Based on these constructions, we first prove that for any adaptive grid and policy, there exists an index j such that the event $A_j = \{t_{j-1} < T_{j-1},\ t_j \geq T_j\}$ happens with sufficiently high probability in world $\mathcal{I}_j$. Then similar to Theorem 5, we prove that in world $\mathcal{I}_j$ there exists a set of problem instances that is difficult to differentiate in the first $j-1$ batches. In addition, event A_j implies that $t_j \geq T_j$, so the worst-case regret is at least $r_j T_j$, which gives the lower bound we need.

The proof of Theorem 6 is deferred to Appendix F. Similar to Corollary 1, we can prove that at least $\Omega(\log\log T)$ rounds of communications are needed to achieve optimal regret bound. This result is formally summarized in Corollary 2.

Corollary 2. *Any Lipschitz bandit algorithm*[3] *needs $\Omega(\log\log T)$ rounds of communications to achieve the optimal regret rate.*

5 Experiments

In this section, we present numerical studies of A-BLiN. In the experiments, we use the arm space $\mathcal{A} = [0,1]^2$ and the expected reward function $\mu(x) = 1 - \frac{1}{2}\|x - x_1\|_2 - \frac{3}{10}\|x - x_2\|_2$, where $x_1 = (0.8,\ 0.7)$ and $x_2 = (0.1,\ 0.1)$. The landscape of μ and the resulting partition is shown in Figure 2(a). As can be seen, the partition is finer in the area closer to the optimal arm $x^* = (0.8,\ 0.7)$.



(a) Partition (b) Regret

Figure 2: Resulting partition and regret of A-BLiN. In Figure 2(a), we show the resulting partition of A-BLiN. The background color denotes the true value of expected reward μ, and blue means high values. The figure shows that the partition is finer for larger values of μ. In Figure 2(b), we show accumulated regret of A-BLiN and zooming algorithm [26]. In the figure, different background colors represent different batches of A-BLiN. For the total time horizon $T = 80000$, A-BLiN needs 4 rounds of communications.

We let the time horizon $T = 80000$, and report the accumulated regret in Figure 2(b). The regret curve is sublinear, which agrees with the regret bound (2). Besides, different background colors in Figure 2(b) represent different batches. For the total time horizon $T = 80000$, A-BLiN only needs 4 rounds of communications. We also present regret curve of zooming algorithm [26] for

[3]including algorithms of which the number of batches can be determined adaptively

comparison. Different from zooming algorithm, regret curve of A-BLiN is approximately piecewise linear, which is because the strategy of BLiN does not change within each batch. Results of more repeated experiments are in Appendix G, as well as experimental results of D-BLiN. Our code is available at `https://github.com/FengYasong-fifol/Batched-Lipschitz-Narrowing`.

6 Conclusion

In this paper, we study Lipschitz bandits with communication constraints, and propose the BLiN algorithm as a solution. We prove that BLiN only need $\mathcal{O}\left(\log\log T\right)$ rounds of communications to achieve the optimal regret rate of best previous Lipschitz bandit algorithms [26, 14] that need T batches. This improvement in number of the batches significantly saves data communication costs. We also provide complexity analysis for this problem. We show that $\Omega(\log\log T)$ rounds of communications are necessary for any algorithm to optimally solve Lipschitz bandit problems. Hence, BLiN algorithm is optimal.

Acknowledgments and Disclosure of Funding

This work was partly supported by the National Key Research and Development Program of China (2020AAA0107600).

References

[1] Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pages 39–75. PMLR, 2017.

[2] Arpit Agarwal, Rohan Ghuge, and Viswanath Nagarajan. Batched dueling bandits. *arXiv preprint arXiv:2202.10660*, 2022.

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

[4] Rajeev Agrawal. Sample mean based index policies by $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

[5] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[6] Patrice Assouad. Plongements Lipschitziens dans $\mathbb{R}^n$. *Bulletin de la Société Mathématique de France*, 111:429–448, 1983.

[7] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[8] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[9] Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.

[10] Dimitris Bertsimas and Adam J Mersereau. A learning approach for interactive marketing to a customer segment. *Operations Research*, 55(6):1120–1135, 2007.

[11] Peter J. Bickel. On some robust estimates of location. *The Annals of Mathematical Statistics*, pages 847–858, 1965.

[12] Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Séminaire de probabilités de Strasbourg*, 12:342–363, 1978.

[13] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[14] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization in $\mathcal{X}$-armed bandits. *Advances in Neural Information Processing Systems*, 22:201–208, 2009.

[15] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. $\mathcal{X}$-armed bandits. *Journal of Machine Learning Research*, 12(5):1655–1695, 2011.

[16] Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the Lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer, 2011.

[17] Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems*, 26:1160–1168, 2013.

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

[19] Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7340–7348, 2021.

[20] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6):1079–1105, 2006.

[21] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32:503–513, 2019.

[22] Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

[23] Kwang-Sung Jun, Kevin Jamieson, Robert Nowak, and Xiaojin Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Artificial Intelligence and Statistics*, pages 139–148. PMLR, 2016.

[24] Nikolai Karpov, Qin Zhang, and Yuan Zhou. Collaborative top distribution identifications with limited interaction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 160–171. IEEE, 2020.

[25] Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 18:697–704, 2005.

[26] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.

[27] Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: smoothing, zooming, and adapting. In *Conference on Learning Theory*, pages 2025–2027. PMLR, 2019.

[28] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[29] Zihan Li and Jonathan Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, pages 92–107. PMLR, 2022.

[30] Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163, 2019.

[31] Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR, 2014.

[32] Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

[33] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.

[34] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[35] Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–87, 2021.

[36] Sudeep Salgia, Sattar Vakili, and Qing Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. In *Advances in Neural Information Processing Systems*, volume 34, pages 28836–28847, 2021.

[37] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[38] Sean Sinclair, Tianyu Wang, Gauri Jain, Siddhartha Banerjee, and Christina Yu. Adaptive discretization for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:3858–3871, 2020.

[39] Aleksandrs Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

[40] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[41] Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 126–146. IEEE, 2019.

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

[43] Tianyu Wang and Cynthia Rudin. Bandits for BMO functions. In *International Conference on Machine Learning*, pages 9996–10006. PMLR, 2020.

[44] Tianyu Wang, Weicheng Ye, Dawei Geng, and Cynthia Rudin. Towards practical lipschitz bandits. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, page 129–138, New York, NY, USA, 2020. Association for Computing Machinery.

[45] Nirandika Wanigasekara and Christina Yu. Nonparametric contextual bandits in an unknown metric space. In *Advances in Neural Information Processing Systems*, volume 32, pages 14684–14694, 2019.

[46] Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. *Advances in Neural Information Processing Systems*, 33:9818–9829, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Lemma 1, Lemma 2 and Lemma 3

Lemma 1. Define

$$\mathcal{E} := \left\{ |\mu(x) - \widehat{\mu}_m(C)| \le r_m + \sqrt{\frac{16 \log T}{n_m}}, \quad \forall 1 \le m \le B_{\text{stop}} - 1,\ \forall C \in \mathcal{A}_m,\ \forall x \in C \right\}.$$

It holds that $\mathbb{P}\left(\mathcal{E}\right) \ge 1 - 2T^{-6}$.

Proof. Fix a cube $C \in \mathcal{A}_m$. Recall the average payoff of cube $C \in \mathcal{A}_m$ is defined as

$$\widehat{\mu}_m(C) = \frac{\sum_{i=1}^{n_m} y_{C,i}}{n_m}.$$

We also have

$$\mathbb{E}\left[\widehat{\mu}_m(C)\right] = \frac{\sum_{i=1}^{n_m} \mu(x_{C,i})}{n_m}.$$

Since $\widehat{\mu}_m(C) - \mathbb{E}\left[\widehat{\mu}_m(C)\right]$ obeys normal distribution $\mathcal{N}\left(0, \frac{1}{n_m}\right)$, Hoeffding inequality gives

$$\mathbb{P}\left(|\widehat{\mu}_m(C) - \mathbb{E}\left[\widehat{\mu}_m(C)\right]| \ge \sqrt{\frac{16 \log T}{n_m}} \right) \le \frac{2}{T^8}.$$

On the other hand, by Lipschitzness of μ, it is obvious that

$$|\mathbb{E}\left[\widehat{\mu}_m(C)\right] - \mu(x)| \le r_m, \quad \forall x \in C.$$

Consequently, we have

$$\mathbb{P}\left(\sup_{x \in C} |\mu(x) - \widehat{\mu}_m(C)| \le r_m + \sqrt{\frac{16 \log T}{n_m}} \right) \ge 1 - \frac{2}{T^8}.$$

For $1 \le m \le B_{\text{stop}} - 1$, the m-th batch is finished, so any cube $C \in \mathcal{A}_m$ is played for not less than 1 time, and thus $|\mathcal{A}_m| \le T$. From here, by similar argument to Lemma F.1 in [38] and Lemma 1 in [30], taking a union bound over $C \in \mathcal{A}_m$ and $1 \le m \le B_{\text{stop}} - 1$ finishes the proof. $\square$

Lemma 2. Under event $\mathcal{E}$, the optimal arm $x^* = \arg\max \mu(x)$ is not eliminated after the first $B_{\text{stop}} - 1$ batches.

Proof. We use C_m^* to denote the cube containing x^* in $\mathcal{A}_m$. Here we proof that C_m^* is not eliminated in round m.

Under event $\mathcal{E}$, for any cube $C \in \mathcal{A}_m$ and $x \in C$, we have

$$\widehat{\mu}(C) - \widehat{\mu}(C_m^*) \le \mu(x) + \sqrt{\frac{16 \log T}{n_m}} + r_m - \mu(x^*) + \sqrt{\frac{16 \log T}{n_m}} + r_m \le 4 r_m.$$

Then from the elimination rule, C_m^* is not eliminated. $\square$

Lemma 3. Under event $\mathcal{E}$, for any $1 \le m \le B_{\text{stop}}$, any $C \in \mathcal{A}_m$ and any $x \in C$, Δ_x satisfies

$$\Delta_x \le 8 r_{m-1}.$$

Proof. For $m = 1$, (3) holds directly from the Lipschitzness of μ. For $m > 1$, let C_{m-1}^* be the cube in $\mathcal{A}_{m-1}$ such that $x^* \in C_{m-1}^*$. From Lemma 2, this cube C_{m-1}^* is well-defined under $\mathcal{E}$. For any cube $C \in \mathcal{A}_m$ and $x \in C$, it is obvious that x is also in the parent of C (the cube in the previous round that contains C), which is denoted by C_{par}. Thus for any $x \in C$, it holds that

$$\Delta_x = \mu^* - \mu(x) \le \widehat{\mu}_{m-1}(C_{m-1}^*) + \sqrt{\frac{16 \log T}{n_{m-1}}} + r_{m-1} - \widehat{\mu}_{m-1}(C_{par}) + \sqrt{\frac{16 \log T}{n_{m-1}}} + r_{m-1},$$

where the inequality uses Lemma 1.

Equality $\sqrt{\frac{16\log T}{n_{m-1}}} = r_{m-1}$ gives that

$$\Delta_x \leq \widehat{\mu}_{m-1}(C^*_{m-1}) - \widehat{\mu}_{m-1}(C_{par}) + 4r_{m-1}.$$

It is obvious that $\widehat{\mu}_{m-1}(C^*_{m-1}) \leq \widehat{\mu}^{\max}_{m-1}$. Moreover, since the cube C_{par} is not eliminated, from the elimination rule we have

$$\widehat{\mu}^{\max}_{m-1} - \widehat{\mu}_{m-1}(C_{par}) \leq 4r_{m-1}.$$

Hence, we conclude that $\Delta_x \leq 8r_{m-1}$. □

B Proof of Theorem 3

Theorem 3. With probability exceeding $1 - \frac{2}{T^6}$, the T-step total regret $R(T)$ of BLiN with Doubling Edge-length Sequence (D-BLiN) satisfies

$$R(T) \leq 528(\log T)^{\frac{1}{d_z+2}} \cdot T^{\frac{d_z+1}{d_z+2}}, \tag{7}$$

where d_z is the zooming dimension of the problem instance. In addition, D-BLiN only needs no more than $\frac{\log T - \log\log T}{d_z+2} + 2$ rounds of communications to achieve this regret rate.

Proof. Since $r_m = \frac{r_{m-1}}{2}$ for Doubling Edge-length Sequence, Lemma 3 implies that every cube $C \in \mathcal{A}_m$ is a subset of $S(16r_m)$. Thus from the definition of zooming number, we have

$$|\mathcal{A}_m| \leq N_{r_m}. \tag{8}$$

Fix any positive number B. Also by Lemma 3, we know that any arm played after batch B incurs a regret bounded by $16r_B$, since the cubes played after batch B have edge length no larger than r_B. Then the total regret occurs after the first B batch is bounded by $16r_BT$.

Thus the regret $R(T)$ can be bounded by

$$R(T) \leq \sum_{m=1}^{B} \sum_{C\in\mathcal{A}_m} \sum_{i=1}^{n_m} \Delta_{x_{C,i}} + 16r_BT, \tag{9}$$

where the first term bounds the regret in the first B batches of D-BLiN, and the second term bounds the regret after the first B batches. If the algorithm stops at batch $\widetilde{B} < B$, we define $\mathcal{A}_m = \varnothing$ for any $\widetilde{B} < m \leq B$ and inequality (9) still holds.

By Lemma 3, we have $\Delta_{C,i} \leq 16r_m$ for all $C \in \mathcal{A}_m$. We can thus bound (9) by

$$\begin{aligned}
R(T) &\leq \sum_{m=1}^{B} |\mathcal{A}_m| \cdot n_m \cdot 16r_m + 16r_BT \\
&\leq \sum_{m=1}^{B} N_{r_m} \cdot n_m \cdot 16r_m + 16r_BT, &(10)\\
&\leq \sum_{m=1}^{B} N_{r_m} \cdot \frac{16\log T}{r_m^2} \cdot 16r_m + 16r_BT, &(11)\\
&= \sum_{m=1}^{B} N_{r_m} \cdot \frac{256\log T}{r_m} + 16r_BT,
\end{aligned}$$

where (10) uses (8), and (11) uses equality $n_m = \frac{16\log T}{r_m}$. Since $r_m = 2^{-m+1}$ and $N_{r_m} \leq r_m^{-d_z} \leq 2^{(m-1)d_z}$, we have

$$R(T) \leq 256 \sum_{m=1}^{B} \frac{2^{(m-1)d_z}\log T}{2^{-(m-1)}} + 16 \cdot 2^{-B+1}T.$$

This inequality holds for any positive B. By choosing $B^* = 1 + \frac{\log \frac{T}{\log T}}{d_z+2}$, we have

$$\begin{aligned} R(T) \leq & 512 \cdot 2^{(B^*-1)(d_z+1)} \log T + 16 \cdot T \cdot 2^{-B^*+1} \\ \leq & 528 T^{\frac{d_z+1}{d_z+2}} (\log T)^{\frac{1}{d_z+2}}. \end{aligned}$$

The above analysis implies that we can achieve the optimal regret rate $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ by letting the *for-loop* run B^* times and finishing the remaining rounds in the *Cleanup* step. In other words, B^*+1 rounds of communications are sufficient for D-BLiN to achieve the regret bound (7). $\square$

C Proof of Remark 2

For rounded ACE Sequence, the quantity $\frac{\widetilde{r}_m}{\widetilde{r}_{m+1}}$ is an integer for any m, so the partition in Line 9 of Algorithm 1 is well-defined. In this section, we show that BLiN with rounded ACE sequence can also achieve the optimal regret bound by using $\mathcal{O}(\log \log T)$ batches. For any m, if there exists $k < m$ such that $\widetilde{r}_m \geq \widetilde{r}_k$, then we skip the m-th batch. It is easy to verify that the following analysis is still valid in this case.

Theorem 7. *With probability exceeding $1 - \frac{2}{T^6}$, the T-step total regret $R(T)$ of Algorithm 1 with rounded ACE sequence satisfies*

$$R(T) \leq \left(\frac{128 C_z \log \log T}{\log \frac{d+2}{d+1-d_z}} + 512 C_z + 8e \right) \cdot T^{\frac{d_z+1}{d_z+2}} (\log T)^{\frac{1}{d_z+2}},$$

where d_z is the zooming dimension of the problem instance. In addition, Algorithm 1 only needs no more than $\frac{2 \log \log T}{\log \frac{d+2}{d+1-d_z}} + 1$ rounds of communications to achieve this regret rate.

Proof. The proof of Theorem 7 is similar to that of Theorem 4.

Firstly, we fix positive number $B^* = \frac{\log \log T - \log(d_z+2)}{\log \frac{d+2}{d+1-d_z}}$ and consider the first $2B^*$ batches. As is summairzed in Remark 2, we bound the regret caused by the first $2B^*$ batches through two different arguments.

For $m = 2k-1$, $1 \leq k \leq B^*$, we have $\widetilde{r}_m = 2^{-\alpha_k}$ and $\widetilde{r}_{m-1} = 2^{-\beta_{k-1}}$, and thus

$$\widetilde{r}_m \geq r_k \quad \text{and} \quad \widetilde{r}_{m-1} \leq r_{k-1}. \tag{12}$$

Let $\mathcal{A}_m^+$ be set of cubes not eliminated in round m. Similar argument to Theorem 4 shows that $|\mathcal{A}_{m-1}^+| \leq C_z \widetilde{r}_{m-1}^{-d_z}$. The total regret of round m is

$$\begin{aligned} R_m &= \sum_{C \in \mathcal{A}_m} \sum_{i=1}^{n_m} \Delta_{x_{C,i}} \\ &\leq |\mathcal{A}_m| \cdot \frac{16 \log T}{\widetilde{r}_m^2} \cdot 8 \widetilde{r}_{m-1} \\ &= \left(\frac{\widetilde{r}_{m-1}}{\widetilde{r}_m} \right)^d |\mathcal{A}_{m-1}^+| \cdot \frac{16 \log T}{\widetilde{r}_m^2} \cdot 8 \widetilde{r}_{m-1} \\ &\leq \left(\frac{\widetilde{r}_{m-1}}{\widetilde{r}_m} \right)^d \cdot C_z \widetilde{r}_{m-1}^{-d_z} \cdot \frac{16 \log T}{\widetilde{r}_m^2} \cdot 8 \widetilde{r}_{m-1} \\ &\leq \widetilde{r}_{m-1}^{d+1-d_z} \cdot \widetilde{r}_m^{-d-2} \cdot 128 C_z \log T \\ &\leq r_{k-1}^{d+1-d_z} \cdot r_k^{-d-2} \cdot 128 C_z \log T \\ &= T^{\frac{d_z+1}{d_z+2}} \cdot 128 C_z \cdot (\log T)^{\frac{1}{d_z+2}}, \end{aligned}$$

where the sixth line follows from (12), and the seventh line follows from (6). Summing over k gives that

$$\sum_{k=1}^{B^*} R_{2k-1} \leq T^{\frac{d_z+1}{d_z+2}} \cdot 128 C_z \cdot B^* \cdot (\log T)^{\frac{1}{d_z+2}} \leq \frac{128 C_z \log \log T}{\log \frac{d+2}{d+1-d_z}} \cdot T^{\frac{d_z+1}{d_z+2}} (\log T)^{\frac{1}{d_z+2}}. \tag{13}$$

For $m = 2k$, $1 \le k \le B^*$, we have $\widetilde{r}_m = 2^{-\beta_k}$ and $\widetilde{r}_{m-1} = 2^{-\alpha_k}$, and thus $\widetilde{r}_m = \frac{1}{2}\widetilde{r}_{m-1}$. Lemma 3 shows that any cube in $\mathcal{A}_m$ is a subset of $S(16\widetilde{r}_m)$, so we have $|\mathcal{A}_m| \le N_{\widetilde{r}_m} \le C_z \widetilde{r}_m^{-d_z}$. Therefore, the total regret of round m is

$$\begin{aligned} R_m &= \sum_{C \in \mathcal{A}_m} \sum_{i=1}^{n_m} \Delta_{x,C_i} \\ &\le |\mathcal{A}_m| \cdot \frac{16 \log T}{\widetilde{r}_m^2} \cdot 16\widetilde{r}_m \\ &\le C_z \widetilde{r}_m^{-d_z} \cdot \frac{16 \log T}{\widetilde{r}_m^2} \cdot 16\widetilde{r}_m \\ &= \widetilde{r}_m^{-d_z - 1} \cdot 256 C_z \log T. \end{aligned}$$

Since $\frac{\widetilde{r}_{2k-2}}{\widetilde{r}_{2k}} = \frac{\widetilde{r}_{2k-2}}{\widetilde{r}_{2k-1}} \cdot \frac{\widetilde{r}_{2k-1}}{\widetilde{r}_{2k}} \ge 2$, summing over k gives that

$$\sum_{k=1}^{B^*} R_{2k} \le \widetilde{r}_{2B^*}^{-d_z - 1} \cdot 512 C_z \log T.$$

The definition of round ACE Sequence shows that

$$\widetilde{r}_{2B^*} = 2^{-\lceil \sum_{i=1}^{B^*} c_i \rceil} = 2^{-\lceil c_1 \left(\frac{1-\eta^{B^*}}{1-\eta} \right) \rceil} = 2^{-\left\lceil \frac{\log \frac{T}{\log T}}{d_z+2} - 1 \right\rceil} \ge \left(\frac{T}{\log T} \right)^{-\frac{1}{d_z+2}},$$

so we have

$$\sum_{k=1}^{B^*} R_{2k} \le \left(\left(\frac{T}{\log T} \right)^{-\frac{1}{d_z+2}} \right)^{-d_z-1} \cdot 512 C_z \log T = T^{\frac{d_z+1}{d_z+2}} \cdot 512 C_z (\log T)^{\frac{1}{d_z+2}}. \tag{14}$$

Similar argument to Theorem 4 shows that the total regret after $2B^*$ batches is upper bounded by $8\widetilde{r}_{2B^*} T$. Since $\widetilde{r}_{2B^*} \le r_{B^*}$, we further have

$$\sum_{m \ge 2B^*} R_m \le 8\widetilde{r}_{2B^*} T \le 8 r_{B^*} T \le 8e \cdot T^{\frac{d_z+1}{d_z+2}} (\log T)^{\frac{1}{d_z+2}}. \tag{15}$$

Combining (13), (14) and (15), we conclude that

$$R(T) \le \left(\frac{128 C_z \log \log T}{\log \frac{d+2}{d+1-d_z}} + 512 C_z + 8e \right) \cdot T^{\frac{d_z+1}{d_z+2}} (\log T)^{\frac{1}{d_z+2}}.$$

The analysis in Theorem 7 implies that we can achieve the optimal regret rate $\widetilde{\mathcal{O}}\left(T^{\frac{d_z+1}{d_z+2}} \right)$ by letting the *for-loop* of Algorithm 1 run $2B^*$ times and finishing the remaining rounds in the *Cleanup* step. In other words, $2B^* + 1$ rounds of communications are sufficient for BLiN to achieve the optimal regret. □

D Proof of Theorem 5

Theorem 5. Consider Lipschitz bandit problems with time horizon T such that the grid of reward communication $\mathcal{T}$ is static and determined before the game. If B rounds of communications are allowed, then for any policy π, there exists a problem instance such that

$$\mathbb{E}[R_T(\pi)] \ge c \cdot (\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} \cdot R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}},$$

where $c > 0$ is a numerical constant independent of B, T, π and $\mathcal{T}$, $R_z(T)$ is defined in (1), and d is the dimension of the arm space.

Proof. To prove Theorem 5, we first show that for any $k > 1$, there exists an instance such that $\mathbb{E}[R_T(\pi)] \geq \frac{t_k}{t_{k-1}^{\frac{1}{d+2}}}$.

Fixing an index $k > 1$, we construct a set of problem instances that is difficult to distinguish. Let $r_k = \frac{1}{t_{k-1}^{\frac{1}{d+2}}}$ and $M_k := t_{k-1} r_k^2 = \frac{1}{r_k^d}$. We can find a set of arms $\mathcal{U}_k = \{u_{k,1}, \cdots, u_{k,M_k}\}$ such that $d_{\mathcal{A}}(u_{k,i}, u_{k,j}) \geq r_k$ for any $i \neq j$. Then we consider a set of problem instances $\mathcal{I}_k = \{I_{k,1}, \cdots, I_{k,M_k}\}$. The expected reward for $I_{k,1}$ is defined as

$$\mu_{k,1}(x) = \begin{cases} \frac{3}{4} r_k, & x = u_{k,1}, \\ \frac{5}{8} r_k, & x = u_{k,j},\ j \neq 1, \\ \max\left\{\frac{r_k}{2}, \max_{u \in \mathcal{U}_k} \{\mu_{k,1}(u) - d_{\mathcal{A}}(x,u)\}\right\}, & otherwise. \end{cases} \tag{16}$$

For $2 \leq i \leq M_k$, the expected reward for $I_{k,i}$ is defined as

$$\mu_{k,i}(x) = \begin{cases} \frac{3}{4} r_k, & x = u_{k,1}, \\ \frac{7}{8} r_k, & x = u_{k,i}, \\ \frac{5}{8} r_k, & x = u_{k,j},\ j \neq 1 \ and \ j \neq i, \\ \max\left\{\frac{r_k}{2}, \max_{u \in \mathcal{U}_k} \{\mu_{k,i}(u) - d_{\mathcal{A}}(x,u)\}\right\}, & otherwise. \end{cases} \tag{17}$$

For all arm pulls in all problem instances, an Gaussian noise sampled from $\mathcal{N}(0,1)$ is added to the observed reward. This noise corruption is independent from all other randomness.

The lower bound of expected regret relies on the following lemma.

Lemma 4. *For any policy π, there exists a problem instance $I \in \mathcal{I}_k$ such that*

$$\mathbb{E}\left[R_T(\pi)\right] \geq \frac{r_k}{32} \cdot \sum_{j=1}^{B} (t_j - t_{j-1}) \exp\left\{-\frac{t_{j-1} r_k^2}{32(M_k - 1)}\right\}.$$

Proof. Let $S_{k,i} = \mathbb{B}(u_{k,i}, \frac{3}{8} r_k)$ (the ball with center $u_{k,i}$ and radius $\frac{3}{8} r_k$). It is easy to verify the following properties of construction (16) and (17):

1. For any $2 \leq i \leq M_k$, $\mu_{k,i}(x) = \mu_{k,1}(x)$ for any $x \in \mathcal{A} - S_{k,i}$;
2. For any $2 \leq i \leq M_k$, $\mu_{k,1}(x) \leq \mu_{k,i}(x) \leq \mu_{k,1}(x) + \frac{r_k}{4}$, for any $x \in S_{k,i}$;
3. For any $1 \leq i \leq M_k$, under $I_{k,i}$, pulling an arm that is not in $S_{k,i}$ incurs a regret at least $\frac{r_k}{8}$.

Let x_t denote the choices of policy π at time t, and y_t denote the reward. Additionally, for $t_{j-1} < t \leq t_j$, we define $\mathbb{P}_{k.i}^t$ as the distribution of sequence $\left(x_1, y_1, \cdots, x_{t_{j-1}}, y_{t_{j-1}}\right)$ under instance $I_{k,i}$ and policy π. It holds that

$$\begin{aligned} \sup_{I \in \mathcal{I}_k} \mathbb{E} R_T(\pi) \geq & \frac{1}{M_k} \sum_{i=1}^{M_k} \mathbb{E}_{\mathbb{P}_{k,i}} \left[R_T(\pi)\right] \\ \geq & \frac{1}{M_k} \sum_{i=1}^{M_k} \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}_{k,i}^t} \left[R^t(\pi)\right] \\ \geq & \frac{r_k}{8} \sum_{t=1}^{T} \frac{1}{M_k} \sum_{i=1}^{M_k} \mathbb{P}_{k,i}^t (x_t \notin S_{k,i}), \end{aligned} \tag{18}$$

where $R^t(\pi)$ denotes the regret incurred by policy π at time t.

From our construction, it is easy to see that $S_{k,j} \cap S_{k,j} = \varnothing$, so we can construct a test Ψ such that $x_t \in S_{k,i}$ implies $\Psi = i$. Then from Lemma 8,

$$\frac{1}{M_k} \sum_{i=1}^{M_k} \mathbb{P}_{k,i}^t(x_t \notin S_{k,i}) \geq \frac{1}{M_k} \sum_{i=1}^{M_k} \mathbb{P}_{k,i}^t(\Psi \neq i) \geq \frac{1}{2M_k} \sum_{i=2}^{M_k} \exp\left\{-D_{KL}\left(\mathbb{P}_{k,1}^t \| \mathbb{P}_{k,i}^t\right)\right\}.$$

Now we calculate $D_{KL}\left(\mathbb{P}_{k,1}^t \| \mathbb{P}_{k,i}^t\right)$. From the chain rule of KL-Divergence, we have

$$\begin{aligned}
D_{KL}\left(\mathbb{P}_{k,1}^t \| \mathbb{P}_{k,i}^t\right) =& D_{KL}\left(\mathbb{P}_{k,1}^t(x_1, y_1, \cdots, x_{t_{j-1}}, y_{t_{j-1}}) \| \mathbb{P}_{k,i}^t(x_1, y_1, \cdots, x_{t_{j-1}}, y_{t_{j-1}})\right) \\
=& D_{KL}\left(\mathbb{P}_{k,1}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1}, x_{t_{j-1}}) \| \mathbb{P}_{k,i}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1}, x_{t_{j-1}})\right) \\
& + \mathbb{E}_{\mathbb{P}_{k,1}}\left(D_{KL}\left(\mathbb{P}_{k,1}^t(y_{t_{j-1}} | x_1, y_1, \cdots, x_{t_{j-1}}) \| \mathbb{P}_{k,i}^t(y_{t_{j-1}} | x_1, y_1, \cdots, x_{t_{j-1}})\right)\right) \\
\leq& D_{KL}\left(\mathbb{P}_{k,1}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1}) \| \mathbb{P}_{k,i}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1})\right) \\
& + \mathbb{E}_{\mathbb{P}_{k,1}}\left(D_{KL}\left(\mathbb{P}_{k,1}^t(y_{t_{j-1}} | x_{t_{j-1}}) \| \mathbb{P}_{k,i}^t(y_{t_{j-1}} | x_{t_{j-1}})\right)\right), \qquad (19) \\
\leq& D_{KL}\left(\mathbb{P}_{k,1}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1}) \| \mathbb{P}_{k,i}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1})\right) \\
& + \mathbb{E}_{\mathbb{P}_{k,1}}\left(D_{KL}\left(N(\mu_{k,1}(x_{t_{j-1}}), 1) \| N(\mu_{k,i}(x_{t_{j-1}}), 1)\right)\right), \qquad (20) \\
=& D_{KL}\left(\mathbb{P}_{k,1}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1}) \| \mathbb{P}_{k,i}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1})\right) \\
& + \mathbb{E}_{\mathbb{P}_{k,1}}\left(\frac{1}{2}\left(\mu_{k,1}(x_{t_{j-1}}) - \mu_{k,i}(x_{t_{j-1}})\right)^2\right), \\
\leq& D_{KL}\left(\mathbb{P}_{k,1}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1}) \| \mathbb{P}_{k,i}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1})\right) \\
& + \mathbb{E}_{\mathbb{P}_{k,1}}\left(\mathbf{1}_{\{x_{t_{j-1}} \in S_{k,i}\}} \cdot \frac{1}{2}\left(\frac{r_k}{4}\right)^2\right) \qquad (21) \\
=& D_{KL}\left(\mathbb{P}_{k,1}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1}) \| \mathbb{P}_{k,i}^t(x_1, y_1, \cdots, x_{t_{j-1}-1}, y_{t_{j-1}-1})\right) \\
& + \frac{r_k^2}{32} \cdot \mathbb{P}_{k,1}\left(x_{t_{j-1}} \in S_{k,i}\right), \qquad (22)
\end{aligned}$$

where (19) uses the non-negativity of KL-Divergence and the conditional independence of the reward, (20) uses that the rewards are corrupted by a standard normal noise, and (21) uses the first two properties of the construction.

From (22), we then decompose the KL-Divergence step by step and conclude that

$$D_{KL}\left(\mathbb{P}_{k,1}^t \| \mathbb{P}_{k,i}^t\right) \leq \frac{r_k^2}{32} \cdot \sum_{s \leq t_{j-1}} \mathbb{P}_{k,1}\left(x_s \in S_{k,i}\right) = \frac{r_k^2}{32} \mathbb{E}_{\mathbb{P}_{k,1}} \tau_i, \qquad (23)$$

where τ_i denotes the number of pulls of arms in $S_{k,i}$ before the batch containing t. Then for all $t \in (t_{j-1}, t_j]$, we have

$$\begin{aligned}
\frac{1}{M_k} \sum_{i=1}^{M_k} \mathbb{P}_{k,i}^t(x_t \notin S_{k,i}) &\geq \frac{1}{2M_k} \sum_{i=2}^{M_k} \exp\left\{-\frac{r_k^2}{32} \mathbb{E}_{\mathbb{P}_{k,1}} \tau_i\right\} \\
&\geq \frac{M_k - 1}{2M_k} \exp\left\{-\frac{r_k^2}{32(M_k - 1)} \sum_{i=2}^{M_k} \mathbb{E}_{\mathbb{P}_{k,1}} \tau_i\right\} \qquad (24) \\
&\geq \frac{1}{4} \exp\left\{-\frac{r_k^2 t_{j-1}}{32(M_k - 1)}\right\}, \qquad (25)
\end{aligned}$$

where (24) uses the Jensen' inequality, and (25) uses the fact that $\sum_{i=2}^{M_k} \tau_i \leq t_{j-1}$. Finally, we substitute (25) to (18) to finish the proof. □

Since $M_k = t_{k-1} r_k^2$, the expected regret of policy π satisfies

$$\begin{aligned}
\mathbb{E}\left[R_T(\pi)\right] &\geq \frac{r_k}{32} \cdot \sum_{j=1}^{B} (t_j - t_{j-1}) \exp\left\{ -\frac{t_{j-1} r_k^2}{32(M_k - 1)} \right\} \\
&\geq \frac{r_k}{32} \cdot \sum_{j=1}^{B} (t_j - t_{j-1}) \exp\left\{ -\frac{t_{j-1} r_k^2}{16 M_k} \right\} \\
&\geq \frac{r_k}{32} \cdot \sum_{j=1}^{B} (t_j - t_{j-1}) \exp\left\{ -\frac{t_{j-1}}{16 t_{k-1}} \right\}
\end{aligned}$$

on instance I defnied in Lemma 4.

By omitting terms with $j > k$ in the above summation, we have

$$\begin{aligned}
\mathbb{E}[R_T(\pi)] &\geq \frac{r_k}{32} \cdot \sum_{j=1}^{B} (t_j - t_{j-1}) \exp\left\{ -\frac{t_{j-1}}{16 t_{k-1}} \right\} \\
&\geq \frac{r_k}{32} \cdot \sum_{j=1}^{k} (t_j - t_{j-1}) \exp\left\{ -\frac{1}{16} \right\} \\
&= \frac{1}{32 e^{\frac{1}{16}}} r_k t_k \\
&= \frac{1}{32 e^{\frac{1}{16}}} \cdot \frac{t_k}{t_{k-1}^{\frac{1}{d+2}}}.
\end{aligned}$$

The above analysis can be applied for any $k > 1$. For the first batch $(0, t_1]$, we can easily construct a set of instances where the worst-case regret is at least t_1, since no information is available during this time. Thus, there exists a problem instance such that

$$\mathbb{E}[R_T(\pi)] \geq \frac{1}{32 e^{\frac{1}{16}}} \max\left\{ t_1, \frac{t_2}{t_1^{\frac{1}{d+2}}}, \cdots, \frac{t_B}{t_{B-1}^{\frac{1}{d+2}}} \right\}.$$

Since $0 < t_1 < \cdots < t_B = T$, we further have

$$\mathbb{E}\left[R_T(\pi)\right] \geq \frac{1}{32 e^{\frac{1}{16}}} \cdot T^{\frac{1 - \frac{1}{d+2}}{1 - \left(\frac{1}{d+2}\right)^B}}. \tag{26}$$

Since $N_r \leq r^{-d}$ holds for all instances,

$$\sum_{r = 2^{-i},\, r \geq r_0} \frac{N_r}{r} \log T \leq 2 r_0^{-d-1} \log T.$$

Then we have

$$\begin{aligned}
R_z(T) =& \inf_{r_0} \left\{ r_0 T + \sum_{r = 2^{-i},\, r \geq r_0} \frac{N_r}{r} \log T \right\} \\
\leq& \inf_{r_0} \left\{ r_0 T + \frac{2}{r_0^{d+1}} \log T \right\} \\
\leq& 2 (\log T)^{\frac{1}{d+2}} T^{1 - \frac{1}{d+2}}.
\end{aligned} \tag{27}$$

As a consequence, for instance I satisfying (26),

$$\begin{aligned}\mathbb{E}[R_T(\pi)] &\geq \frac{1}{32e^{\frac{1}{16}}}\left(\frac{1}{2(\log T)^{\frac{1}{d+2}}}\right)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}} \cdot R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}} \\ &\geq \frac{1}{32e^{\frac{1}{16}} \cdot 2^{\frac{1}{1-\frac{1}{d+2}}}} \cdot (\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} \cdot R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}} \\ &\geq \frac{1}{128e^{\frac{1}{16}}} \cdot (\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} \cdot R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}}.\end{aligned}$$

Hence, the proof is completed. □

E Proof of Corollary 1

Corollary 1. Any algorithm needs $\Omega(\log\log T)$ rounds of communications to optimally solve Lipschitz bandit problems, for the case that the times of reward communication are predetermined and static.

Proof. From (26), the expected regret is lower bounded by $\frac{1}{32e^{\frac{1}{16}}} \cdot T^{\frac{1-\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}}$. When B is sufficiently large, the bound becomes $\frac{1}{32e^{\frac{1}{16}}} \cdot T^{1-\frac{1}{d+2}}$. Here we seek for the minimum B such that

$$\frac{\frac{1}{32e^{\frac{1}{16}}} \cdot T^{\frac{1-\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}}}{\frac{1}{32e^{\frac{1}{16}}} \cdot T^{1-\frac{1}{d+2}}} \leq C \tag{28}$$

for some constant C.

Calculation shows that

$$\frac{\frac{1}{32e^{\frac{1}{16}}} \cdot T^{\frac{1-\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}}}{\frac{1}{32e^{\frac{1}{16}}} \cdot T^{1-\frac{1}{d+2}}} = \left(T^{\frac{d+1}{d+2}}\right)^{\frac{1}{(d+2)^B-1}}. \tag{29}$$

Substituting (29) to (28) and taking log on both sides yield that

$$\frac{d+1}{d+2} \cdot \frac{\log T}{(d+2)^B - 1} \leq \log C$$

and

$$(d+2)^B \geq \frac{d+1}{(d+2)\log C} \cdot \log T + 1.$$

Taking log on both sides again yields that

$$B \geq \frac{\log\left[\left(\frac{d+1}{(d+2)\log C}\right)\log T + 1\right]}{\log(d+2)}.$$

Therefore, $\Omega(\log\log T)$ rounds of communications are necessary for any algorithm to optimally solve Lipschitz bandit problems. □

F Proof of Theorem 6

Theorem 6. Consider Lipschitz bandit problems with time horizon T such that the grid of reward communication $\mathcal{T}$ is adaptively determined by the player. If B rounds of communications are allowed, then for any policy π, there exists an instance such that

$$\mathbb{E}\left[R_T(\pi)\right] \geq c\frac{1}{B^2}(\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}},$$

where $c > 0$ is a numerical constant independent of B, T, π and $\mathcal{T}$, $R_z(T)$ is defined in (1), and d is the dimension of the arm space.

Proof. Firstly, we define $r_j = \frac{1}{T_{j-1}^{\varepsilon}B}$ and $M_j = \frac{1}{r_j^d}$. From the definition, we have

$$T_{j-1}r_j^2 = \frac{1}{r_j^d B^2} = \frac{M_j}{B^2}. \tag{30}$$

For $1 \leq j \leq B$, we can find sets of arms $\mathcal{U}_j = \{u_{j,1}, \cdots, u_{j,M_j}\}$ such that (a) $d_{\mathcal{A}}(u_{j,m}, u_{j,n}) \geq r_j$ for any $m \neq n$, and (b) $u_{1,M_1} = \cdots = u_{B,M_B}$.

Then we present the construction of worlds $\mathcal{I}_1, \cdots, \mathcal{I}_B$. For $1 \leq j \leq B-1$, we let $\mathcal{I}_j = \{I_{j,k}\}_{k=1}^{M_j-1}$, and the expected reward of $I_{j,k}$ is defined as

$$\mu_{j,k}(x) = \begin{cases} \frac{r_1}{2} + \frac{r_j}{16} + \frac{r_B}{16}, & x = u_{j,k}, \\ \frac{r_1}{2} + \frac{r_B}{16}, & x = u_{j,M_j}, \end{cases} \tag{31}$$

and $\mu_{j,k}(x) = \max\left\{\frac{r_1}{2}, \max_{u \in \mathcal{U}_j}\left\{\mu_{j,k}(u) - d_{\mathcal{A}}(x,u)\right\}\right\}$, otherwise. For $j = B$, we let $\mathcal{I}_B = \{I_B\}$. The expected reward of I_B is defined as

$$\mu_B(u_{B,M_B}) = \frac{r_1}{2} + \frac{r_B}{16} \tag{32}$$

and $\mu_B(x) = \max\left\{\frac{r_1}{2}, \mu_B(u_{B,M_B}) - d_{\mathcal{A}}(x, u_{B,M_B})\right\}$, otherwise.

As mentioned above, based on these constructions, we first show that for any adaptive grid $\mathcal{T} = \{t_0, \cdots, t_B\}$, there exists an index j such that $(t_{j-1}, t_j]$ is sufficiently large in world $\mathcal{I}_j$. More formally, for each $j \in [B]$, and event $A_j = \{t_{j-1} < T_{j-1},\ t_j \geq T_j\}$, we define the quantities $p_j := \frac{1}{M_j-1}\sum_{k=1}^{M_j-1}\mathbb{P}_{j,k}(A_j)$ for $j \leq B-1$ and $p_B := \mathbb{P}_B(A_B)$, where $\mathbb{P}_{j,k}(A_j)$ denotes the probability of the event A_j under instance $I_{j,k}$ and policy π. For these quantities, we have the following lemma.

Lemma 5. *For any adaptive grid $\mathcal{T}$ and policy π, it holds that $\sum_{j=1}^{B} p_j \geq \frac{7}{8}$.*

Proof. For $1 \leq j \leq B-1$ and $1 \leq k \leq M_j - 1$, we define $S_{j,k} = \mathbb{B}(u_{j,k}, \frac{3}{8}r_j)$, which is the ball centered as $u_{j,k}$ with radius $\frac{3}{8}r_j$. It is easy to verify the following properties of our construction (31) and (32):

1. $\mu_{j,k}(x) = \mu_B(x)$ for any $x \notin S_{j,k}$;
2. $\mu_B(x) \leq \mu_{j,k}(x) \leq \mu_B(x) + \frac{r_j}{8}$, for any $x \in S_{j,k}$.

Let x_t denote the choices of policy π at time t, and y_t denote the reward. For $t_{j-1} < t \leq t_j$, we define $\mathbb{P}_{j,k}^t$ (resp. $\mathbb{P}_B^t$) as the distribution of sequence $\left(x_1, y_1, \cdots, x_{t_{j-1}}, y_{t_{j-1}}\right)$ under instance $I_{j,k}$ (resp. I_B) and policy π. Since event A_j can be completely described by the observations up to time T_{j-1} (A_j is an event in the σ-algebra where $\mathbb{P}_{j,k}^{T_{j-1}}$ and $\mathbb{P}_B^{T_{j-1}}$ are defined on), we can use the definition of total variation to get

$$|\mathbb{P}_B(A_j) - \mathbb{P}_{j,k}(A_j)| = |\mathbb{P}_B^{T_{j-1}}(A_j) - \mathbb{P}_{j,k}^{T_{j-1}}(A_j)| \leq TV\left(\mathbb{P}_B^{T_{j-1}}, \mathbb{P}_{j,k}^{T_{j-1}}\right).$$

For the total variation, we apply Lemma 7 to get

$$\frac{1}{M_j-1}\sum_{k=1}^{M_j-1} TV\left(\mathbb{P}_B^{T_{j-1}}, \mathbb{P}_{j,k}^{T_{j-1}}\right) \le \frac{1}{M_j-1}\sum_{k=1}^{M_j-1}\sqrt{1-\exp\left(-D_{KL}\left(\mathbb{P}_B^{T_{j-1}}\|\mathbb{P}_{j,k}^{T_{j-1}}\right)\right)}.$$

An argument similar to (23) yields that

$$D_{KL}\left(\mathbb{P}_B^{T_{j-1}}\|\mathbb{P}_{j,k}^{T_{j-1}}\right) \le \frac{r_j^2}{128}\mathbb{E}_{\mathbb{P}_B}\tau_k,$$

where τ_k denotes the number of pulls which is in $S_{j,k}$ before the batch containing T_{j-1}. Combining the above two inequalities gives

$$\begin{aligned}
\frac{1}{M_j-1}\sum_{k=1}^{M_j-1} TV\left(\mathbb{P}_B^{T_{j-1}}, \mathbb{P}_{j,k}^{T_{j-1}}\right) &\le \frac{1}{M_j-1}\sum_{k=1}^{M_j-1}\sqrt{1-\exp\left(-\frac{r_j^2}{128}\mathbb{E}_{\mathbb{P}_B}\tau_k\right)} \\
&\le \sqrt{1-\exp\left(-\frac{r_j^2}{128(M_j-1)}\mathbb{E}_{\mathbb{P}_B}\left[\sum_{k=1}^{M_j-1}\tau_k\right]\right)} \qquad (33)\\
&\le \sqrt{1-\exp\left(-\frac{r_j^2 T_{j-1}}{128(M_j-1)}\right)} \qquad (34)\\
&\le \sqrt{1-\exp\left(-\frac{1}{64B^2}\right)} \qquad (35)\\
&\le \frac{1}{8B},
\end{aligned}$$

where (33) uses Jensen's inequality, (34) uses the fact that $\sum_{k=1}^{M_j-1}\tau_k \le T_{j-1}$, and (35) uses (30).

Plugging the above results implies that

$$|\mathbb{P}_B(A_j)-p_j| \le \frac{1}{M_j-1}\sum_{k=1}^{M_j-1}|\mathbb{P}_B(A_j)-\mathbb{P}_{j,k}(A_j)| \le \frac{1}{8B}.$$

Since $\sum_{j=1}^{B}\mathbb{P}\left(A_j\right) \ge \mathbb{P}\left(\cup_{j=1}^{B}A_j\right)=1$, it holds that

$$\sum_{j=1}^{B} p_j \ge \mathbb{P}_B(A_M) + \sum_{j=1}^{B-1}\left(\mathbb{P}_B(A_j)-\frac{1}{8B}\right) \ge \frac{7}{8}. \qquad \square$$

Lemma 5 implies that there exists some j such that $p_j > \frac{7}{8B}$. Then similar to Theorem 5, we show that the worst-case regret of the policy in world $\mathcal{I}_j$ gives the lower bound we need.

Lemma 6. *For adaptive grid $\mathcal{T}$ and policy π, if index j satisfies $p_j \ge \frac{7}{8B}$, then there exists a problem instance I such that $\mathbb{E}\left[R_T(\pi)\right] \ge c\frac{1}{B^2}(\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}}$, where $c>0$ is a numerical constant independent of B, T, π and $\mathcal{T}$.*

Proof. Here we proceed with the case where $j \le B-1$. The case for $j=B$ can be proved analogously.

For any $1 \le k \le M_j - 1$, we construct a set of problem instances $\mathcal{I}_{j,k} = \left(I_{j,k,l}\right)_{1\le l\le M_j}$. For $l \ne k$, the expected reward of $I_{j,k,l}$ is defined as

$$\mu_{j,k,l}(x) = \begin{cases} \mu_{j,k}(x) + \frac{3r_j}{16}, & x = u_{j,l}, \\ \mu_{j,k}(x), & x \in \mathcal{U}_j \quad \text{and} \quad x \ne u_{j,l}, \\ \max\left\{\frac{r_1}{2}, \max_{u\in\mathcal{U}_j}\left\{\mu_{j,k,l}(u) - d_{\mathcal{A}}(x,u)\right\}\right\}, & otherwise, \end{cases}$$

where $\mu_{j,k}$ is defined in (31). For $l = k$, we let $\mu_{j,k,k} = \mu_{j,k}$.

We define $C_{j,k} = \mathbb{B}\left(u_{j,k}, \frac{r_j}{4}\right)$, and our construction $\mathcal{I}_{j,k}$ has the following properties:

1. For any $l \neq k$, $\mu_{j,k,l}(x) = \mu_{j,k,k}(x)$ for any $x \notin C_{j,l}$;
2. For any $l \neq k$, $\mu_{j,k,k}(x) \leq \mu_{j,k,l}(x) \leq \mu_{j,k,k}(x) + \frac{3r_j}{16}$ for any $x \in C_{j,l}$;
3. For any $1 \leq l \leq M_j$, under $I_{j,k,l}$, pulling an arm that is not in $C_{j,l}$ incurs a regret at least $\frac{r_j}{16}$.

Let x_t denote the choices of policy π at time t, and y_t denote the reward. For $t_{j-1} < t \leq t_j$, we define $\mathbb{P}^t_{j,k,l}$ as the distribution of sequence $\left(x_1, y_1, \cdots, x_{t_{j-1}}, y_{t_{j-1}}\right)$ under instance $I_{i,j,k}$ and policy π. From similar argument in (18), it holds that

$$\sup_{I \in \mathcal{I}_{j,k}} \mathbb{E}\left[R_T(\pi)\right] \geq \frac{r_j}{16} \sum_{t=1}^{T} \frac{1}{M_j} \sum_{l=1}^{M_j} \mathbb{P}^t_{j,k,l}(x_t \notin C_{j,l}). \tag{36}$$

From our construction, it is easy to see that $C_{j,k_1} \cap C_{j,k_2} = \varnothing$ for any $k_1 \neq k_2$, so we can construct a test Ψ such that $x_t \notin C_{j,k}$ implies $\Psi \neq k$. By Lemma 8 with a star graph on $[K]$ with center k, we have

$$\frac{1}{M_j} \sum_{l=1}^{M_j} \mathbb{P}^t_{j,k,l}(x_t \notin C_{j,l}) \geq \frac{1}{M_j} \sum_{l \neq k} \int \min\left\{d\mathbb{P}^t_{j,k,k}, d\mathbb{P}^t_{j,k,l}\right\}. \tag{37}$$

Combining (36) and (37) gives

$$\begin{aligned}
\sup_{I \in \mathcal{I}_{j,k}} \mathbb{E}\left[R_T(\pi)\right] &\geq \frac{r_j}{16} \sum_{t=1}^{T} \frac{1}{M_j} \sum_{l \neq k} \int \min\left\{d\mathbb{P}^t_{j,k,k}, d\mathbb{P}^t_{j,k,l}\right\} \\
&\geq \frac{r_j}{16} \sum_{t=1}^{T_j} \frac{1}{M_j} \sum_{l \neq k} \int \min\left\{d\mathbb{P}^t_{j,k,k}, d\mathbb{P}^t_{j,k,l}\right\} \\
&\geq \frac{r_j T_j}{16} \cdot \frac{1}{M_j} \sum_{l \neq k} \int \min\left\{d\mathbb{P}^{T_j}_{j,k,k}, d\mathbb{P}^{T_j}_{j,k,l}\right\} \qquad (38) \\
&\geq \frac{r_j T_j}{16} \cdot \frac{1}{M_j} \sum_{l \neq k} \int_{A_j} \min\left\{d\mathbb{P}^{T_j}_{j,k,k}, d\mathbb{P}^{T_j}_{j,k,l}\right\} \qquad (39) \\
&\geq \frac{r_j T_j}{16} \cdot \frac{1}{M_j} \sum_{l \neq k} \int_{A_j} \min\left\{d\mathbb{P}^{T_{j-1}}_{j,k,k}, d\mathbb{P}^{T_{j-1}}_{j,k,l}\right\}, \qquad (40)
\end{aligned}$$

where (38) follows from data processing inequality of total variation and the equation $\int \min\{dP, dQ\} = 1 - TV(P, Q)$, (39) restricts the integration to event A_j, and (40) holds because the observations at time T_j are the same as those at time T_{j-1} under event A_j.

For the term $\int_{A_j} \min\left\{d\mathbb{P}^{T_{j-1}}_{j,k,k}, d\mathbb{P}^{T_{j-1}}_{j,k,l}\right\}$, it holds that

$$\begin{aligned}
\int_{A_j} \min\left\{d\mathbb{P}^{T_{j-1}}_{j,k,k}, d\mathbb{P}^{T_{j-1}}_{j,k,l}\right\} &= \int_{A_j} \frac{d\mathbb{P}^{T_{j-1}}_{j,k,k} + d\mathbb{P}^{T_{j-1}}_{j,k,l} - \left|d\mathbb{P}^{T_{j-1}}_{j,k,k} - d\mathbb{P}^{T_{j-1}}_{j,k,l}\right|}{2} \\
&= \frac{\mathbb{P}^{T_{j-1}}_{j,k,k}(A_j) + \mathbb{P}^{T_{j-1}}_{j,k,l}(A_j)}{2} - \frac{1}{2} \int_{A_j} \left|d\mathbb{P}^{T_{j-1}}_{j,k,k} - d\mathbb{P}^{T_{j-1}}_{j,k,l}\right| \\
&\geq \left(\mathbb{P}^{T_{j-1}}_{j,k,k}(A_j) - \frac{1}{2} TV\left(\mathbb{P}^{T_{j-1}}_{j,k,k}, \mathbb{P}^{T_{j-1}}_{j,k,l}\right)\right) - TV\left(\mathbb{P}^{T_{j-1}}_{j,k,k}, \mathbb{P}^{T_{j-1}}_{j,k,l}\right) \qquad (41) \\
&= \mathbb{P}_{j,k}(A_j) - \frac{3}{2} TV\left(\mathbb{P}^{T_{j-1}}_{j,k,k}, \mathbb{P}^{T_{j-1}}_{j,k,l}\right), \qquad (42)
\end{aligned}$$

where (41) uses the inequality $|\mathbb{P}(A)-\mathbb{Q}(A)|\le TV(\mathbb{P},\mathbb{Q})$, and (42) holds because $I_{j,k}=I_{j,k,k}$ and A_j can be determined by the observations up to T_{j-1}.

We use an argument similar to (23) to get

$$D_{KL}\left(\mathbb{P}_{j,k,k}^{T_{j-1}}\|\mathbb{P}_{j,k,l}^{T_{j-1}}\right)\le\frac{1}{2}\cdot\left(\frac{3r_j}{16}\right)^2\mathbb{E}_{\mathbb{P}_{j,k}}\tau_l\le\frac{r_j^2}{32}\mathbb{E}_{\mathbb{P}_{j,k}}\tau_l,$$

where τ_l denotes the number of pulls which is in $S_{j,l}$ before the batch of time T_{j-1}. Then from Lemma 7, we have

$$\begin{aligned}
\frac{1}{M_j}\sum_{l\neq k}TV\left(\mathbb{P}_{j,k,k}^{T_{j-1}},\mathbb{P}_{j,k,l}^{T_{j-1}}\right)&\le\frac{1}{M_j}\sum_{l\neq k}\sqrt{1-\exp\left(-D_{KL}\left(\mathbb{P}_{j,k,k}^{T_{j-1}}\|\mathbb{P}_{j,k,l}^{T_{j-1}}\right)\right)}\\
&\le\frac{1}{M_j}\sum_{l\neq k}\sqrt{1-\exp\left(-\frac{r_j^2}{32}\mathbb{E}_{\mathbb{P}_{j,k}}\tau_l\right)}\\
&\le\frac{M_j-1}{M_j}\sqrt{1-\exp\left(-\frac{r_j^2}{32(M_j-1)}\sum_{l\neq k}\mathbb{E}_{\mathbb{P}_{j,k}}\tau_l\right)}\\
&\le\frac{M_j-1}{M_j}\sqrt{1-\exp\left(-\frac{r_j^2T_{j-1}}{32(M_j-1)}\right)}\\
&\le\frac{M_j-1}{M_j}\sqrt{1-\exp\left(-\frac{M_j}{32(M_j-1)B^2}\right)} \qquad (43)\\
&\le\frac{M_j-1}{M_j}\sqrt{\frac{M_j}{32(M_j-1)B^2}}\\
&\le\frac{1}{4B}, \qquad (44)
\end{aligned}$$

where (43) uses (30).

Combining (40), (42) and (44) yields that

$$\begin{aligned}
\sup_{I\in\mathcal{I}_{j,k}}\mathbb{E}\left[R_T(\pi)\right]&\ge\frac{1}{16}r_jT_j\left(\frac{\mathbb{P}_{j,k}(A_j)}{2}-\frac{3}{8B}\right)\\
&\ge\frac{1}{16B}T^{\frac{1-\varepsilon}{1-\varepsilon^B}}\left(\frac{\mathbb{P}_{j,k}(A_j)}{2}-\frac{3}{8B}\right).
\end{aligned}$$

This inequality holds for any $k\le M_j-1$. Averaging over k yields

$$\begin{aligned}
\sup_{I\in\cup_{k\le M_j-1}\mathcal{I}_{j,k}}\mathbb{E}\left[R_T(\pi)\right]&\ge\frac{1}{16B}T^{\frac{1-\varepsilon}{1-\varepsilon^B}}\left(\frac{1}{2(M_j-1)}\sum_{k=1}^{M_j-1}\mathbb{P}_{j,k}(A_j)-\frac{3}{8B}\right)\\
&\ge\frac{1}{16B}T^{\frac{1-\varepsilon}{1-\varepsilon^B}}\left(\frac{7}{16B}-\frac{3}{8B}\right) \qquad (45)\\
&\ge\frac{1}{256B^2}T^{\frac{1-\varepsilon}{1-\varepsilon^B}},
\end{aligned}$$

where the second inequality holds from $p_j\ge\frac{7}{8B}$.

Consequently, combining (45) and (27) yields that

$$\begin{aligned}
\sup_{I \in \cup_{k \leq M_j - 1} \mathcal{I}_{j,k}} \mathbb{E}\left[R_T(\pi)\right] &\geq \frac{1}{256B^2} T^{\frac{1-\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} \\
&\geq \frac{1}{256 \cdot 2^{\frac{1}{1-\frac{1}{d+2}}}} \frac{1}{B^2} (\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}} \\
&\geq \frac{1}{1024} \cdot \frac{1}{B^2} (\log T)^{-\frac{\frac{1}{d+2}}{1-\left(\frac{1}{d+2}\right)^B}} R_z(T)^{\frac{1}{1-\left(\frac{1}{d+2}\right)^B}},
\end{aligned}$$

which finishes the proof. □

Finally, combining the above two lemmas, we arrive at the lower bound in Theorem 6. □

G Experimental results

G.1 Repeated experiments of A-BLiN

We present results of A-BLiN with some random seeds below (the curve of zooming algorithm in Figure 2(b) in the paper is the average of 10 repeated experiments), where the figure legends and labels are the same as whose in Figure 2(b). These results stably agree with the plot in the paper. The reason we did not present averaged regret curve of A-BLiN in Figure 2(b) is that we want to show the batch pattern of a single A-BLiN run in the figure. Averaging across different runs breaks the batch pattern. As an example, one stochastic run may end the first batch after 100 observations, while another may end the third batch after 110 observations.

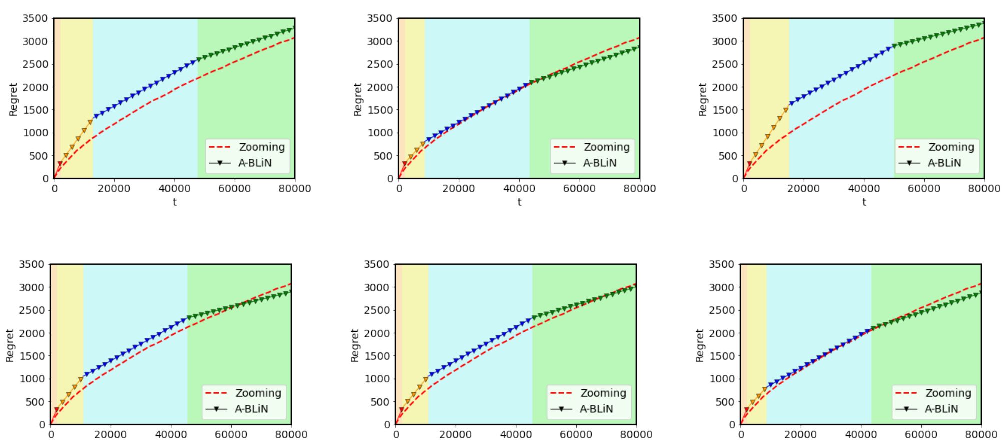


Figure 3: Results of A-BLiN with some random seeds. The figure legends and labels are the same as whose in Figure 2(b).

G.2 Experimental results of D-BLiN

We run D-BLiN to solve the same problem in Section 5. The partition and elimination process of this experiment is presented in Figure 4, which shows that the optimal arm x^* is not eliminated during the game, and only 6 rounds of communications are needed for time horizon $T = 80000$. Moreover, we present the resulting partition and the accumulated regret in Figure 5.

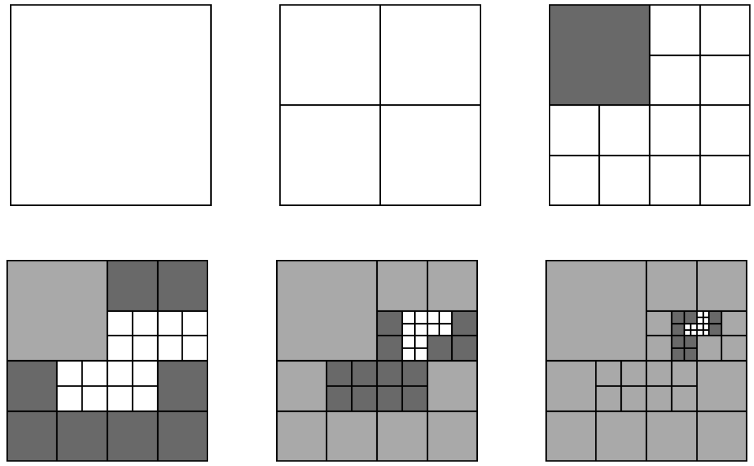
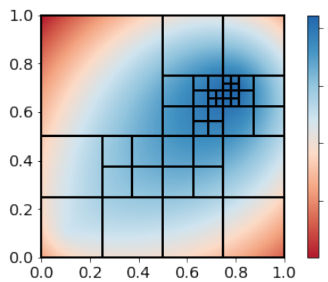
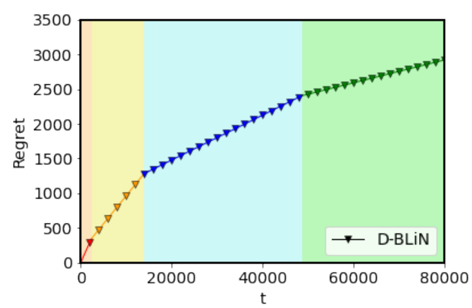


Figure 4: Partition and elimination process of D-BLiN. The i-th subfigure shows the pattern before the i-th batch. Dark gray cubes are those eliminated in the most recent batch, while the light gray ones are those eliminated in earlier batches. For the total time horizon $T = 80000$, D-BLiN needs 6 rounds of communications.



(a) Partition



(b) Regret

Figure 5: Resulting partition and regret of D-BLiN. In Figure 5(a), we show the resulting partition of D-BLiN. The background color denotes the true value of expected reward μ, and blue means high values. The figure shows that the partition is finer for larger values of μ. In Figure 5(b), we show accumulated regret of D-BLiN. In the figure, different background colors represent different batches. For the total time horizon $T = 80000$, D-BLiN needs 6 rounds of communications (the first two batches are too small and are combined with the third batch in the plot).

H Auxiliary Technical Tools

Lemma 7 (Bretagnolle-Huber Inequality[12]). *Let P and Q be any probability measures on the same probability space. It holds that*

$$TV(P,Q) \leq \sqrt{1 - \exp\left(-D_{KL}(P\|Q)\right)} \leq 1 - \frac{1}{2}\exp\left(-D_{KL}(P\|Q)\right).$$

Lemma 8 ([21]). *Let $Q_1, \cdots, Q_n$ be probability measures over a common probability space $(\Omega, \mathcal{F})$, and $\Psi : \Omega \to [n]$ be any measurable function (i.e., test). Then for any tree $T = ([n], E)$ with vertex set $[n]$ and edge set E, we have*

1. $\frac{1}{n}\sum_{i=1}^{n} Q_i(\Psi \neq i) \geq \frac{1}{n}\sum_{(i,j)\in E} \int \min\{dQ_i, dQ_j\}$;
2. $\frac{1}{n}\sum_{i=1}^{n} Q_i(\Psi \neq i) \geq \frac{1}{2n}\sum_{(i,j)\in E} \exp\left(-D_{KL}(Q_i\|Q_j)\right)$.