

New Directions in Bandit Learning: Singularities and Random
Walk Feedback

by

Tianyu Wang

Department of Computer Science
Duke University

Date: _____

Approved:

Cynthia Rudin, Advisor

Cynthia Rudin

Xiuyuan Cheng

Rong Ge

Alexander Volfovsky

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computer Science
in the Graduate School of
Duke University

2021

ABSTRACT

NEW DIRECTIONS IN BANDIT LEARNING:
SINGULARITIES AND RANDOM WALK FEEDBACK

by

Tianyu Wang

Department of Computer Science
Duke University

Date: _____

Approved:

Cynthia Rudin, Advisor

Cynthia Rudin

Xiuyuan Cheng

Rong Ge

Alexander Volfovsky

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Computer Science
in the Graduate School of
Duke University

2021

Copyright © 2021 by Tianyu Wang
All rights reserved

Abstract

My thesis focuses new directions in bandit learning problems. In Chapter 1, I give an overview of the bandit learning literature, which lays the discussion framework for studies in Chapters 2 and 3. In Chapter 2, I study bandit learning problem in metric measure spaces. I start with multi-armed bandit problem with Lipschitz reward, and propose a practical algorithm that can utilize greedy tree training methods and adapts to the landscape of the reward function. In particular, the study provides a Bayesian perspective to this problem. Also, I study bandit learning for Bounded Mean Oscillation (BMO) functions, where the goal is to “maximize” a function that may go to infinity in parts of the space. For an unknown BMO function, I will present algorithms that efficiently finds regions with high function values. To handle possible singularities and unboundedness in BMO functions, I will introduce the new notion of δ -regret – the difference between the function values along the trajectory and a point that is optimal after removing a δ -sized portion of the space. I will show that my algorithm has $\mathcal{O}\left(\frac{\kappa \log T}{T}\right)$ average T -step δ -regret, where κ depends on δ and adapts to the landscape of the underlying reward function.

In Chapter 3, I will study bandit learning with random walk trajectories as feedback. In domains including online advertisement and social networks, user behaviors can be modeled as a random walk over a network. To this end, we study a novel bandit learning problem, where each arm is the starting node of a random walk in a network and the reward is the length of the walk. We provide a comprehensive understanding of this formulation

by studying both the stochastic and the adversarial setting. In the stochastic setting, we observe that, there exists a difficult problem instance on which the following two seemingly conflicting facts simultaneously hold: 1. No algorithm can achieve a regret bound independent of problem intrinsic information theoretically; and 2. There exists an algorithm whose performance is independent of problem intrinsic in terms of tail of mistakes. This reveals an intriguing phenomenon in general semi-bandit feedback learning problems. In the adversarial setting, we establish novel algorithms that achieve regret bound of order $\tilde{O}(\sqrt{\kappa T})$, where κ is a constant that depends on the structure of the graph, instead of number of arms (nodes). This bound significantly improves regular bandit algorithms, whose complexity depends on number of arms (nodes).

Contents

Abstract	iv
List of Figures	x
List of Tables	xi
Acknowledgements	xii
1 Introduction	1
2 Bandit Learning in Metric Spaces	7
2.1 Lipschitz Bandits: A Bayesian Approach	7
2.1.1 Introduction	7
2.1.2 Main Results: TreeUCB Framework and a Bayesian Perspective	9
2.1.3 Empirical Study	29
2.1.4 Conclusion	31
2.2 Bandits for BMO Functions	32
2.2.1 Introduction	32
2.2.2 Preliminaries	34
2.2.3 Problem Setting: BMO Bandits	36
2.2.4 Solve BMO Bandits via Partitioning	38
2.2.5 Achieve Poly-log Regret via Zooming	44

2.2.6	Experiments	51
2.2.7	Conclusion	52
3	Bandit Learning with Random Walk Feedback	53
3.1	Introduction	53
3.1.1	Related Works	56
3.2	Problem Setting	58
3.3	Stochastic Setting	59
3.3.1	Reduction to Standard MAB	60
3.3.2	Is this Problem Much Easier than Standard MAB?	62
3.3.3	Regret Analysis	70
3.4	Adversarial Setting	71
3.4.1	Analysis of Algorithm 6	74
3.4.2	Lower Bound for the Adversarial Setting	79
3.5	Experiments	81
3.6	Conclusion	82
4	Conclusion	83
	Appendices	85
A	Supplementary Materials for Chapter 2	86

A.1	Proof of Lemma 2	86
A.2	Proof of Lemma 3	87
A.2.1	Proof of Proposition 5	91
A.2.2	Proof of Proposition 6	92
A.3	Proof of Lemma 4	92
A.4	Proof of Theorem 1	93
A.5	Proof of Proposition 2	94
A.6	Proof of Proposition 3	95
A.7	Elaboration of Remark 6	96
A.8	Proof of Theorem 3	99
B	Supplementary Materials for Chapter 3	103
B.1	Additional Details for the Stochastic Setting	103
B.1.1	Concentrations of Estimators	103
B.1.2	Proof of Theorem 9	105
B.1.3	Proof of Theorem 10	108
B.1.4	Greedy Algorithm for the Stochastic Setting	113
B.2	Proofs for the Adversarial Setting	116
B.2.1	Proof of Theorem 12	121
B.2.2	Additional Propositions	126

Bibliography	128
Biography	142

List of Figures

1.1	A bandit octopus.	2
2.1	Example reward function (in color gradient) with an example partitioning.	10
2.2	The left subfigure is the metric learned by Algorithm 1 (2.9). The right subfigure is the smoothed version of this learned metric.	29
2.3	The estimates for a function with respect to a given partition.	30
2.4	Performance of TUCB against benchmark methods in tuning neural networks.	31
2.5	Graph of $f(x) = -\log(x)$, with δ and f^δ annotated. This function is an unbounded BMO function.	37
2.6	Example of terminal cubes, pre-parent and parent cubes.	46
2.7	Algorithms 3 and 4 on Himmelblau’s function (left) and Styblinski–Tang function (right).	51
2.8	Landscapes of test functions used in Section 2.1.3. Left: (Rescaled) Himmelblau’s function. Right: (Rescaled) Styblinski-Tang function.	52
3.1	Problem instances constructed to prove Theorem 8. The edge labels denote edge transition probabilities in $\mathfrak{J}/\mathfrak{J}'$	66
3.2	A plot of function $f(x) = \frac{1-\sqrt{x}}{1+\sqrt{x}}$, $x \in [0, 1]$. This shows that in Theorem 11, the dependence on graph connectivity is highly non-linear.	74
3.3	Experimental results for Chapter 3.5.	80
3.4	The network structure for experiments in Chapter 3.5.	82

List of Tables

2.1	Settings for the SVHN experiments.	25
2.2	Settings for CIFAR-10 experiments.	26

Acknowledgements

When I started my PhD study, I knew little about the journey ahead. There were multiple points where I almost failed. After five years of adventures, how much I have transformed!

I am grateful to my advisor, Prof. Cynthia Rudin, for her guidance and advise. Lessons I learnt from Prof. Rudin are not only technical, but also philosophical. She has taught me how to define problems, how to collaborate, how to write academic papers and give talks. I am still practicing and improving skills learnt from her.

I would like to thank Duke University and the Department of Computer Science. Administratively and financially, the university and the department have supported me through my PhD study. Also, a thank you to the Alfred P. Sloan Foundation for supporting me via the Duke Energy Data Analytics Fellowship.

I would like to thank all my co-authors during my PhD study. They are, alphabetically, M. Usaid Awan, Siddhartha Banerjee, Dawei Geng, Gauri Jain, Yameng Liu, Marco Morucci, Sudeepa Roy, Cynthia Rudin, Sean Sinclair, Alexander Volfovsky, Zizhuo Wang, Lin Yang, Weicheng Ye, Christina Lee Yu.

I would like to thank all the course instructors and teachers. The techniques and methodologies I learnt from them are invaluable. I also thank Duke University and the Department of Computer Science for providing rich learning resources.

Very importantly, I am grateful to my parents.

Looking ahead into the future, I will maintain a high standard for myself, to live up to the names of Duke University, the Department of Computer Science, my advisor Cynthia Rudin, and all my collaborators.

Chapter 1

Introduction

Bandit learning algorithms seek to answer the following critical question in sequential decision making:

In interacting with the environment, when to exploit the historically good options, and when to explore the decision space?

This intriguing question, and the corresponding exploitation-exploration tension arise in many, if not all, online decision making problems. Bandit learning algorithms find applications ranging from experiment design [Rob52] to online advertising [LCLS10].

In the classic bandit learning problem, an agent is interacting with an unknown and possibly changing environment. This agent has a set of choices (called arms in bandit community), and is trying to maximize the total reward, while learning the environment. The performance is usually measured by regret. The regret is defined as the total difference, summed over time, between the agent's choices, and a hindsight optimal option. We seek to design algorithms with a sub-linear regret in time. This ensures that when we can run the algorithm long enough, we are often choosing the best options.

Three Different Settings

In this part, I classify bandit problems into three different settings, based on how the environment may change. I survey these three settings, with focus on their early stage development. While there are many other possible ways to classify bandit problems (e.g., based on whether the feedback is full-information, semi-bandit, or bandit), I hope this classification communicates my first hunch of dissecting a bandit problem.

Stochastic Bandit. Stochastic bandit is perhaps the oldest form of bandit learning, as it can date its history back to about 90 years ago. [Tho33]. In this setting, the agent is

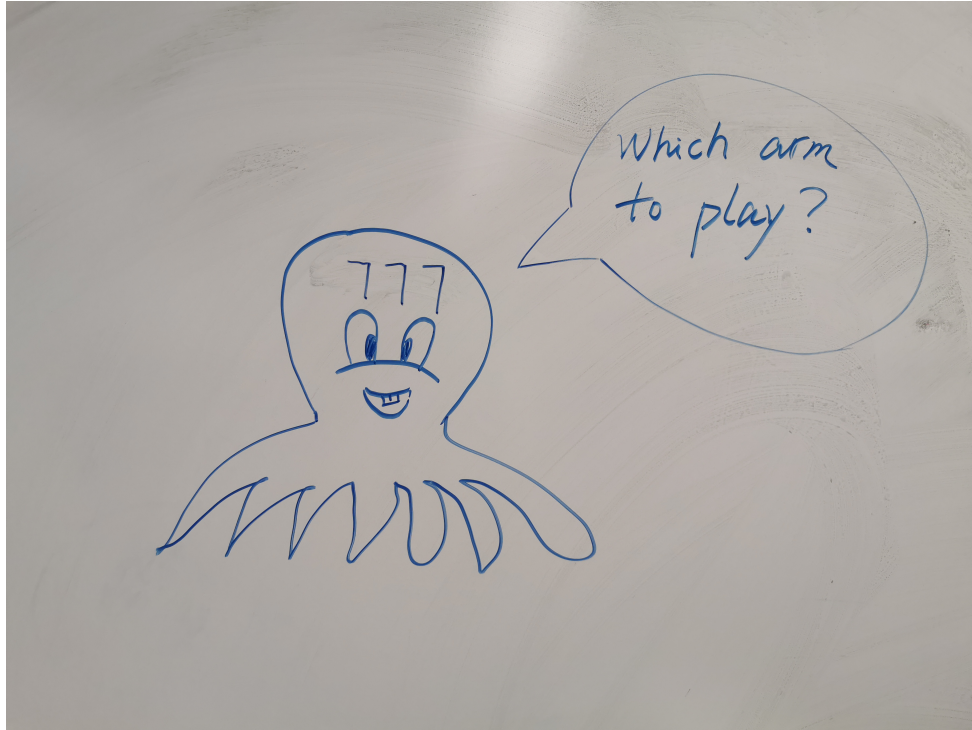


Figure 1.1: A bandit octopus.

faced with K arms, each with an unknown and unchanging distribution. The agent's goal is to find out the arm with best expected reward while interacting with this environment. Perhaps the oldest algorithm for stochastic bandit is Thompson Sampling [Tho33], which maintains a belief distribution for each option, and makes decisions according to samples from the belief distributions. It is worth noting that the tight performance guarantee for this elegant method remains unknown until 2011 [AG13]. This paper by Thompson, together with the seminal papers [Rob52, LR85], lays the foundation of modern stochastic bandit. In their seminal papers [Rob52, LR85], the ideas of certainty, confidence, and Upper Confidence Bound (UCB) were introduced, and the asymptotic lower bound for all bandit learning algorithms was derived. The UCB method plays arms that maximize the empirical mean estimate plus a confidence interval of the estimate. The UCB algorithms are pushed to its modern form later by [Agr95b, ACBF02]. This UCB principle has inspired a long list of works: linear bandits (e.g., [Aue02, AYPS11, LCLS10]), stochastic combinatorial bandit

(e.g., [CWY13, KWAS15]), Lipschitz bandits (e.g., [Kle05, KSU08, BMSS11]), just to name a few.

Markov Bandit. In the 70s, a different formulation parallel to the stochastic bandit problem was invented, Gittins studied a Markov bandit problem where each arm is associated with a state that can evolve [Git79]. An arm’s state evolves in a Markovian fashion, and only changes when it is pulled. Each time, the agent receives a reward that depends on the arm pulled and the state of the pulled arm. The goal is to maximize the total long-term reward (with respect to some discount factor). The upfront motivation in the original paper was job scheduling. Each time the worker selects a job to work on. The status of the selected job is then updated, and a reward is given to the worker. This formulation later finds especially useful for medical trials – each time a patient receives treatment, her internal state evolves. Naturally, this Markov formulation later expands into the works of restless bandits [Whi88], where the states may evolve even if the corresponding arms are not pulled.

Adversarial Bandit. Another line of work parallel to the above two settings, is that of adversarial bandits. In an adversarial bandit problem, the agent plays arms and an adversary alters rewards. No restriction except for boundedness is imposed on the adversarial rewards. Perhaps such adversarial bandits are captured in casinos for real, as the seminal paper’s title “Gambling in a Rigged Casino” suggests [ACBFS95]. In such case, the performance metric is usually oblivious regret, or e.g., dynamic regret (e.g., [HW13]). The oblivious regret is the difference between the total reward of agent’s choices, and the total reward of a hindsight-optimal arm. Most adversarial algorithms are some form of exponential weight algorithm [ACBFS95, FS97]. In these algorithms, a Boltzmann’s softmax is applied, so that historically more rewarding arms are more likely to be played in the future. What’s fascinating about exponential weights is the intrinsic connection to mirror descent, follow the leader, and large body of works in online learning and optimization therein (e.g., [SS⁺11]). Quite intriguingly, the adversarial bandit problem also relates to zeroth-order convex optimization, other than in the mirror descent sense. In particular,

Flexman et al. [FKM05] studied a zeroth-order optimization method for convex functions. They leverages the Stokes' Theorem, and use a single-point zeroth-order information to estimate the first order gradient.

Recent Development and New Directions¹

Recently, the field of bandit learning has expanded tremendously. In particular, recent study of bandit learning relates to optimization in multiple ways. The most natural connection is through the problem of pure exploration and the notion of simple regret [BMS09, ABM10, BMS11, KDO⁺16a, SCY19]. In such problems, the agent ignores the need to exploit, and simply seeks to find the best option. In such problems, the goal is to minimize simple regret, which is the difference between reward of the best arm of the reward of a single-step choice. When the underlying reward is endowed with convexity, the problem naturally has intriguing connection to optimization. A general regime is to use Stokes' theorem to estimate gradient using zeroth-order information (or bandit feedback) [FKM05]. Variations with specific geometric structure (e.g., [HL14]) and projection-free versions (e.g., [CZK19]) have also been developed. The lower bound has also been derived using a clever global construction [Sha13]. An intrinsic connection between bandit algorithms and optimization is: exponential weights algorithms can be viewed as mirror descent with negated entropy as the mirror map. This means that the multiplicative weight update is equivalently a gradient step in the dual space. Works down this line inherits rich relations to online learning and optimization in general; See e.g., [SS⁺11, B⁺15, H⁺16] for an overview of this connection.

Recent study of bandit learning also relates to reinforcement learning. Many bandit learning principles have been used in reinforcement learning, especially recently, after the successful usage of the Upper Confidence Tree method [KS06] in solving GO games [SHM⁺16]. Practitioners also often perturb the policy with the ϵ -greedy principle when the action space is large [SB98]. Recently, both the UCB mechanism and the exponential weight methods are used in reinforcement learning. The use of UCB in reinforcement

¹This part serves as a general picture before I introduce the new directions studied in this thesis. More contextualized discussions of related works will be presented in Chapters 2 and 3.

learning is a natural extension to Q -learning. In Q -learning, one maximizes the estimated Q function to decide an action. Combined with the UCB principle, one maximizes the upper confidence bound of the estimated Q -function to learn a policy. Jin et al use the UCB principle to design adaptive exploration strategies in finite-horizon tabular Markov decision processes [JAZBJ18]. This was later extended to infinite-horizon case [WDCW19]. Under this optimistic principle, Linear [YW20, JYWJ20] and Lipschitz [SWJ⁺20] variations are also studied. The exponential weights methods are also used in Markov decision processes. To use exponential weights in this setting, a general recipe is to set the probability of playing action a in state x being proportional to the exponentiated Q -value $Q(x, a)$ [EDKM05, NGS10, AYBB⁺19, JLL⁺20].

In addition to studies related to optimization and reinforcement learning, many other variations have been studied. To name a few, combinatorial bandits have been investigated in both stochastic [CWY13] and adversarial setting [CBL12]. Bandit with switching cost studies a setting where playing different arms in two consecutive rounds incurs an additional cost [AHT88, DDKP14]. Also, bandits with knapsack provides a general framework for bandit learning problems with resource constraints [BKS13, AD16, ISSS19].

In this thesis, I will focus on two problems in bandit learning. I will study new direction and new use case of bandit learning. In Chapter 2, I study bandit learning problem in metric measure spaces. I start with multi-armed bandit problem with Lipschitz reward, and propose a practical algorithm that can utilize greedy tree training methods and adapts to the landscape of the reward function. In particular, the study provides a Bayesian perspective to this problem. The connection to finite horizon Lipschitz reinforcement learning is also discussed. Also, I study bandit learning for Bounded Mean Oscillation (BMO) functions, where the goal is to “maximize” a function that may go to infinity in parts of the space. For an unknown BMO function, I will present algorithms that efficiently finds regions with high function values. To handle possible singularities and unboundedness in BMO functions, I will introduce the new notion of δ -regret – the difference between the function values along the trajectory and a point that is optimal after removing a δ -sized portion of the space. I

will show that my algorithm has $\mathcal{O}\left(\frac{\kappa \log T}{T}\right)$ average T -step δ -regret, where κ depends on δ and adapts to the landscape of the underlying reward function. In Chapter 3, I will study work motivated by online advertising. Nowadays, millions of people open mobile apps, and randomly browse items in the app. The browsing over items in the app can be modeled as a random walk over the items. To this end, I propose a Markov random walk model to capture the dynamics of browsing behavior. The items (e.g., video clips) are modeled as nodes in a graph. Each epoch t , a user arrives at an entrance item – opens the app, and performs a random walk over the graph of items – every click on an item is a transition to a new node. The user closes the app when the random walk hits an absorbing node. This model leads to an important real-world question: How to make good recommendations in this model? I will provide an algorithmic answer to this question in different settings, along with theoretical guarantees.

Chapter 2

Bandit Learning in Metric Spaces

2.1 Lipschitz Bandits: A Bayesian Approach

2.1.1 Introduction

A stochastic bandit problem assumes that payoffs are noisy and are drawn from an unchanging distribution. The study of stochastic bandit problems started with the discrete arm setting, where the agent is faced with a finite set of choices. Classic works on this problem include Thompson sampling [Tho33, AG12], Gittins index [Git79], ϵ -greedy strategies [SB98], and upper confidence bound (UCB) methods [LR85, ACBF02]. One recent line of work on stochastic bandit problems considers the case where the arm space is infinite. In this setting, the arms are usually assumed to be in a subset of the Euclidean space (or a more general metric space), and the expected payoff function is assumed to be a function of the arms. Some works along this line model the expected payoff as a linear function of the arms [Aue02, DHK08, LCLS10, AYPS11, AG13]; some algorithms model the expected payoff as Gaussian processes over the arms [SKKS10a, CPV14, dFSZ12]; some algorithms assume that the expected payoff is a Lipschitz function of the arms [Sli14, KSU08, BMSS11, MCP14]; and some assume locally Hölder payoffs on the real line [AOS07]. When the arms are continuous and equipped with a metric, and the expected payoff is Lipschitz continuous in the arm space, we refer to the problem as a stochastic Lipschitz bandit problem. In addition, when the agent's decisions are made with the aid of contextual information, we refer to the problem as a contextual stochastic Lipschitz bandit problem. Not many works [BMSS11, KSU08, MCP14] have considered the general Lipschitz bandit problem without making strong assumptions on the smoothness of rewards in context-arm space. In this part, we focus our study on this general (contextual) stochastic Lipschitz

bandit problem, and provide practical algorithms for use in data science applications.

Specifically, we propose a framework that converts a general decision tree algorithm into an algorithm for stochastic Lipschitz bandit problems. We use a novel analysis that links our algorithms to Gaussian processes; though the underlying rewards do not need to be generated by any Gaussian process. Based on this connection, we can use a novel hierarchical Bayesian model to design a new (UCB) index. This new index solves two main problems suffered by partition based bandit algorithms. Namely, (1) within each bin of the partition, all arms are treated the same; (2) disjoint bins do not use information from each other.

Empirically, we show that using adaptively learned partitions, Lipschitz bandit algorithms can be used for hard real-world problems such as hyperparameter tuning for neural networks.

Relation to prior work: One general way of solving stochastic Lipschitz bandit problems is to finely discretize (partition) the arm space and treat the problem as a finite-arm problem. An Upper Confidence Bound (UCB) strategy can thus be used. Previous algorithms of this kind include the `UniformMesh` algorithm [KSU08], the HOO algorithm [BMSS11], and the (contextual) Zooming Bandit algorithm [KSU08, Sli14]. While all these algorithms employ different analysis techniques, we show that as long as a discretization of the arm space fulfills certain requirements (outlined in Theorem 1), these algorithms (or a possibly modified version) can be analyzed in a unified framework.

The practical problem with previous methods is that they require either a fine discretization of the full arm space or restrictive control of the partition formation (e.g., Zooming rule [KSU08]), leading to implementations that are not flexible. By fitting decision trees that are grown adaptively during the run of the algorithm, our partition can be learned from data. This advantage enables the algorithm to outperform leading methods for Lipschitz bandits (e.g. [BMSS11, KSU08]) and for zeroth order optimization (e.g. [MC14, LJD⁺16]) on hard real-world problems that can involve difficult arm space and reward landscape. As shown in the experiments, in neural network hyperparameter tuning, our methods can

outperform the state-of-the-art benchmark packages that are tailored for hyperparameter selection.

In summary, our contributions are: **1)** We develop a novel stochastic Lipschitz bandit framework, TreeUCB and its contextual counterpart Contextual TreeUCB. Our framework converts a general decision tree algorithm into a stochastic Lipschitz bandit algorithm. Algorithms arising from this framework empirically outperform benchmarks methods. **2)** We develop a new analysis framework, which can be used to recover previous known bounds, and design a new principled acquisition function in bandits and zero-th order optimization.

2.1.2 Main Results: TreeUCB Framework and a Bayesian Perspective

The TreeUCB framework

Stochastic bandit algorithms, in an online fashion, explore the decision space while exploit seemingly good options. The performance of the algorithm is typically measured by regret. In this part, we focus our study on the following setting. A payoff function is defined over an arm space that is a compact doubling metric space (\mathcal{A}, d) , the payoff function of interest is $f : \mathcal{A} \rightarrow [0, 1]$, and the actual observations are given by $y(a) = f(a) + \epsilon_a$. In our setting, the noise distribution ϵ_a could vary with a , as long as it is uniformly mean zero, almost surely bounded, and independent of f for every a . Our results easily generalize to sub-Gaussian noise [Sha11]. In the analysis, we assume that the (expected) payoff function f is Lipschitz in the sense that $\forall a, a' \in \mathcal{A}, |f(a) - f(a')| \leq Ld(a, a')$ for some Lipschitz constant L . An agent is interacting with this environment in the following fashion. At each round t , based on past observations $(a_1, y_1, \dots, a_{t-1}, y_{t-1})$, 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 an agent executing algorithm Alg, the regret incurred up to time

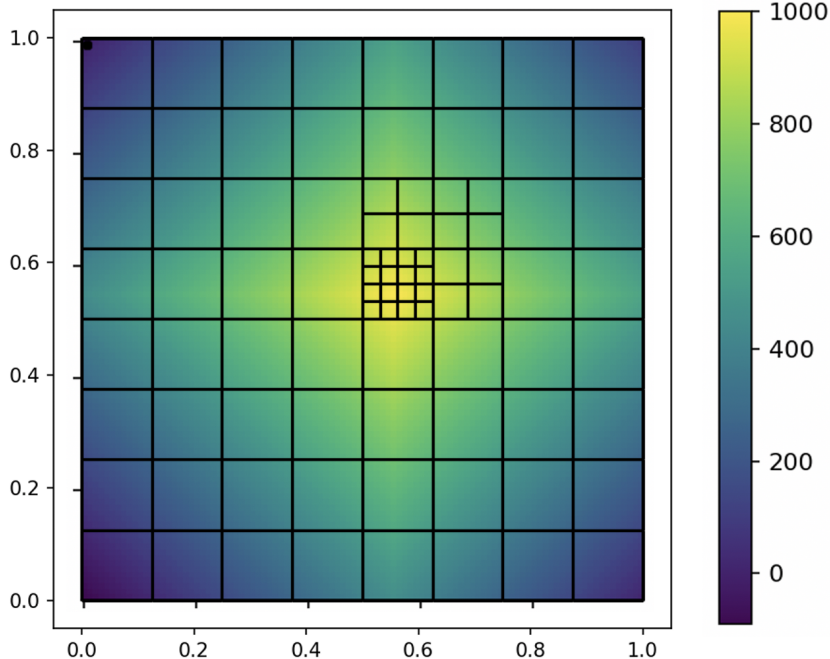


Figure 2.1: Example reward function (in color gradient) with an example partitioning.

T is defined to be:

$$R_T(\text{Alg}) = \sum_{t=1}^T (f(a^*) - f(a_t)),$$

where a^* is the global maximizer of f .

Any TreeUCB algorithm runs by maintaining a sequence of finite partitions of the arm space. Intuitively, at each step t , TreeUCB treats the problem as a finite-arm bandit problem with respect to the partition bins at t , and chooses an arm uniformly at random within the chosen bin. The partition bins become smaller and smaller as the algorithm runs. Thus, at any time t , we maintain a partition $\mathcal{P}_t = \{P_t^{(1)}, \dots, P_t^{(k_t)}\}$ of the input space. That is, $P_t^{(1)}, \dots, P_t^{(k_t)}$ are subsets of \mathcal{A} , are mutually disjoint and $\cup_{i=1}^{k_t} P_t^{(i)} = \mathcal{A}$.

As an example, Figure 2.1 shows an partitioning of the input space, with the underlying reward function shown by color gradient. In an algorithm run, we collect data and estimate the reward with respect to the partition. Based on the estimate, we select a “box” to play next.

Each element in the partition is called a region and by convention $\mathcal{P}_0 = \{\mathcal{A}\}$. The regions could be leaves in a tree, or chosen in some other way.

Given any t , if for any $P^{(i)} \in \mathcal{P}_{t+1}$, there exists $P^{(j)} \in \mathcal{P}_t$ such that $P^{(i)} \subset P^{(j)}$, we say that $\{\mathcal{P}_t\}_{t \geq 0}$ is a sequence of **nested partitions**. In words, at round t , some regions (or no regions) of the partition are split into multiple regions to form the partition at round $t + 1$. We also say that the partition **grows finer**.

Based on the partition \mathcal{P}_t at time t , we define an auxiliary function – the Region Selection function.

Definition 1 (Region Selection Function). *Given partition \mathcal{P}_t , function $p_t : \mathcal{A} \rightarrow \mathcal{P}_t$ is called a Region Selection Function with respect to \mathcal{P}_t if for any $a \in \mathcal{A}$, $p_t(a)$ is the region in \mathcal{P}_t containing a .*

As the name TreeUCB suggests, our framework follows an Upper Confidence Bound (UCB) strategy. In order to define our Upper Confidence Bound, we require several definitions.

Definition 2. *Let \mathcal{P}_t be the partition of \mathcal{A} at time t ($t \geq 1$) and let p_t be the Region Selection Function associated with \mathcal{P}_t . Let $(a_1, y_1, a_2, y_2, \dots, a_{t'}, y_{t'})$ be the observations received up to time t' ($t' \geq 1$). We define*

- the count function $n_{t,t'}^0 : \mathcal{A} \rightarrow \mathbb{R}$, such that

$$n_{t,t'}^0(x) = \sum_{i=1}^{t'} \mathbb{I}[x_i \in p_t(x)].$$

- the corrected average function $m_{t,t'} : \mathcal{A} \rightarrow \mathbb{R}$, such that

$$m_{t,t'}(a) = \begin{cases} \frac{\sum_{i=1}^{t'} y_i \mathbb{I}[a_i \in p_t(a)]}{n_{t,t'}^0(a)}, & \text{if } n_{t,t'}^0(a) > 0; \\ 1, & \text{otherwise.} \end{cases} \quad (2.1)$$

- the corrected count function, such that

$$n_{t,t'}(x) = \max(1, n_{t,t'}^0(x)). \quad (2.2)$$

When $t = t'$, we shorten the notation from $m_{t,t'}$ to m_t , $n_{t,t'}^0$ to n_t^0 , and $n_{t,t'}$ to n_t .

In words, $n_{t,t'}^0(a)$ is the number of points among $(a_1, a_2, \dots, a_{t'})$ that are in the same region as arm a , with regions as elements in \mathcal{P}_t . We also denote by $D(\mathcal{S})$ the diameter of $\mathcal{S} \subset \mathcal{A}$, and $D(\mathcal{S}) := \sup_{a', a'' \in \mathcal{S}} d(a', a'')$.

At time t , based on the partition and observations, our bandit algorithm uses, for $a \in \mathcal{A}$

$$U_t(a) = m_{t-1}(a) + C \sqrt{\frac{4 \log t}{n_{t-1}(a)}} + M \cdot D(p_t(a)), \quad (2.3)$$

for some C and M as the Upper Confidence Bound of arm a ; and we play an arm with the highest U_t value (with ties broken uniformly at random).

Remark 1. *As we will discuss in Section 2.1.2, the upper confidence index for our decision can take different forms other than (2.3).*

Here C depends on the almost sure bound on the reward, and M depends on the Lipschitz constant of the expected reward, which are both problem intrinsics.

Since U_t is a piece-wise constant function in the arm-space and is constant within each region, playing an arm with the highest U_t with random tie-breaking is equivalent to selecting the best region (under UCB) and randomly selecting an arm within the region. After deciding which arm to play, we update the partition into a finer one if eligible. This strategy, TreeUCB, is summarized in Algorithm 1. We also provide a provable guarantee for TreeUCB algorithms in Theorem 1.

Theorem 1. *Suppose that the payoff function f defined on a compact domain \mathcal{A} satisfies $f(a) \in [0, 1]$ for all a and is Lipschitz. Let \mathcal{P}_t be the partition at time t in Algorithm 1. If the tree fitting rule \mathcal{R} satisfies*

- (1) $\{\mathcal{P}_t\}_{t \geq 0}$ is a sequence of nested partitions (or the partition grows finer);
- (2) $|\mathcal{P}_t| = o(t^\gamma)$ for some $\gamma < 1$;
- (3) $D(p_t(a)) = o(1)$ for all $a \in \mathcal{A}$, where

$$D(p_t(a)) := \sup_{a', a'' \in p_t(a)} d(a', a'')$$

is the diameter of region $p_t(a)$;

- (4) given all realized observations $\{(a_t, y_t)\}_{t=1}^T$, the partitions $\{\mathcal{P}_t\}_{t=1}^T$ are deterministic;

Algorithm 1 TreeUCB (TUCB)

1: Parameter: $M \geq 0$ ($M \geq L$). $C > 0$. Tree fitting rule \mathcal{R} that satisfies 1–4 in Theorem 1.

/** C depends on the a.s. bound of the reward./

/** M depends on the Lipschitz constant of the expected reward./

2: **for** $t = 1, 2, \dots, T$ **do**

3: Fit the tree f_{t-1} using rule \mathcal{R} on observations $(a_1, y_1, a_2, y_2, \dots, a_{t-1}, y_{t-1})$.

4: With respect to the partition \mathcal{P}_{t-1} defined by leaves of f_{t-1} , define m_{t-1}, n_{t-1} as in (2.1) and (2.2). Play

$$a_t \in \arg \max_{a \in \mathcal{A}} \{U_t(a)\}, \quad (2.4)$$

where U_t is defined in (2.3). Ties are broken uniformly at random.

5: Observe the reward y_t .

then the regret for Algorithm 1 satisfies

$$\lim_{T \rightarrow \infty} \frac{R_T(\text{TUCB})}{T} = 0$$

with probability 1.

The above assumptions are all mild and reasonable. For item 1, we can use incremental tree learning [Utg89] to enforce nested partitions. For item 2, we may put a cap (that may depend on t) on the depth of the tree to constrain it. For item 3, we may put a cap (that may depend on t) on tree leaf diameters to ensure it. For item 4, any non-random tree learning rule meets this criteria, since in this case, the randomness only comes from the data (and/or number of data points observed).

We now discuss the proof of Theorem 1. Throughout the rest of the paper, we use $\tilde{\mathcal{O}}$ to omit constants and poly-log terms unless otherwise noted. To prove Theorem 1, we first use Claims 1 and 2 to bound the single step regret, we then use Lemma 1 and Assumptions (1) – (3) to bound the total regret.

To start with, we first present the following two claims, which may also be carefully extracted from previous works (e.g., [BMSS11]).

Claim 1. *For an arbitrary arm a , and time t , with probability at least $1 - \frac{1}{t^4}$, we have,*

$$|m_{t-1}(a) - f(a)| \leq L \cdot D(p_{t-1}(a)) + C \sqrt{\frac{4 \log t}{n_{t-1}(a)}}$$

for a constant C that depends only on the a.s. bound of the reward.

Claim 2. *At any t , with probability at least $1 - \frac{1}{t^4}$, the single step regret satisfies:*

$$f(a^*) - f(a_t) \leq 2L \cdot D(p_{t-1}(a_t)) + 2C \sqrt{\frac{4 \log t}{n_{t-1}(a_t)}} \quad (2.5)$$

for a constant C , that depends only on the a.s. bound of the reward.

In Section 2.1.2, we prove general versions of Claims 1 and 2.

As the tree (partition) grows finer, the term $n_{t-1}(a)$ is not necessarily increasing with t (for an arbitrary fixed a). Therefore part of the difficulty is in bounding $\sum_{t=1}^T \frac{1}{n_{t-1}(a_t)}$. Next, we introduce a new set of inequalities, which we call “point scattering” inequalities in Lemma 1 to bound this term.

Lemma 1 (Point Scattering Inequalities). *For an arbitrary sequence of points a_1, a_2, \dots in a space \mathcal{A} , and any sequence of nested partitions $\mathcal{P}_1, \mathcal{P}_2, \dots$ of the same space \mathcal{A} , we have, for any T ,*

$$\sum_{t=1}^T \frac{1}{n_{t-1}(a_t)} \leq e^{|\mathcal{P}_T|} \log \left(1 + (e-1) \frac{T}{|\mathcal{P}_T|} \right), \quad (2.6)$$

$$\sum_{t=1}^T \frac{1}{1 + n_{t-1}^0(a_t)} \leq |\mathcal{P}_T| \left(1 + \log \frac{T}{|\mathcal{P}_T|} \right), \quad (2.7)$$

$$\sum_{t=1}^T \left(\frac{1}{1 + n_{t-1}^0(a_t)} \right)^\alpha \leq \frac{1}{1-\alpha} |\mathcal{P}_T|^{\alpha T^{1-\alpha}}, \quad 0 < \alpha < 1, \quad (2.8)$$

where n_{t-1}^0 and n_{t-1} are the count and corrected count function as in Definition 2, and $|\mathcal{P}_T|$ is the cardinality of the finite partition \mathcal{P}_T .

As defined in Definition 2, $n_{t-1}^0(a_t)$ is the number of points that are in the same bin (in partition \mathcal{P}_{t-1}) as a_t . Also, $n_{t-1}(a_t)$ is the “corrected” version of $n_{t-1}^0(a_t)$: $n_{t-1}(a_t) = \max(1, n_{t-1}^0(a_t))$.

Remark 2. *We shall notice that (2.6) allows us to somewhat “look one step ahead of time”, since it uses the values $\{n_{t-1}(a_t)\}_t$ - the corrected counts without including a_t . This is because n_{t-1} is computed using points up to time $t - 1$. The equation (2.7) is different from (2.6) in the sense that $\{1 + n_{t-1}^0(a_t)\}_t$ are essentially the counts including a_t . While, with proper modification, both (2.6) and (2.7) can be used to derive Theorem 1, we shall not ignore the difference between (2.6) and (2.7).*

Proof of (2.6)

We use a novel constructive trick to derive (2.6). This trick and the usefulness of the result (Remarks 2 and 3 and Section 2.1.2) mark our major technical contribution. The trick is to consider the incidence matrix of which points are within the same partition bin, and use this matrix as if it were a covariance matrix for a Gaussian process. Then, we use knowledge about Gaussian processes to bound the sum of the inverse of the number of points in each bin over time.

For each T , we construct a hypothetical noisy degenerate Gaussian process. We are not assuming our payoffs are drawn from these Gaussian processes. We only use these Gaussian processes as a proof tool. To construct these noisy degenerate Gaussian processes, we define the kernel functions $k_T : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$,

$$k_T(a, a') = \begin{cases} 1, & \text{if } p_T(a) = p_T(a') \\ 0, & \text{otherwise.} \end{cases} \quad (2.9)$$

where p_T is the region selection function defined with respect to \mathcal{P}_T . The kernel k_T is positive semi-definite as shown in Proposition 1.

Proposition 1. *The kernel defined in (2.9) is positive semi-definite for any $T \geq 1$.*

Proof. For any x_1, \dots, x_n in where the kernel $k_T(\cdot, \cdot)$ is defined, the Gram matrix $K = [k_T(x_i, x_j)]_{n \times n}$ can be written into block diagonal form where diagonal blocks are all one matrices and off-diagonal blocks are all zeros with proper permutations of rows and columns. Thus without loss of generality, for any vector $\mathbf{v} = [v_1, v_2, \dots, v_n] \in \mathbb{R}^n$, $\mathbf{v}^\top K \mathbf{v} = \sum_{b=1}^B \left(\sum_{j:i_j \text{ in block } b} v_{i_j} \right)^2 \geq 0$ where the first summation is taken over all diagonal blocks and B is the total number of diagonal blocks in the Gram matrix. \square

Now, at any time T , let us consider the model $\tilde{y}(a) = g(a) + e_T$ where g is drawn from a Gaussian process $g \sim \mathcal{GP}(0, k_T(\cdot, \cdot))$ and $e_T \sim \mathcal{N}(0, s_T^2)$. Suppose that the arms and hypothetical payoffs $\{(a_1, \tilde{y}_1), (a_2, \tilde{y}_2), \dots, (a_t, \tilde{y}_t)\}$ are observed from this Gaussian process. The posterior variance for this Gaussian process after the observations at a_1, a_2, \dots, a_t is

$$\sigma_{T,t}^2(a) = k_T(a, a) - \mathbf{k}_a^\top (K + s_T^2 I)^{-1} \mathbf{k}_a,$$

where $\mathbf{k}_a = [k_T(a, a_1), \dots, k_T(a, a_t)]^\top$, $K = [k_T(a_i, a_j)]_{t \times t}$ and I is the identity matrix. In other words, $\sigma_{T,t}^2(a)$ is the posterior variance using points up to time t with the kernel defined by the partition at time T . After some matrix manipulation, we know that

$$\sigma_{T,t}^2(a) = 1 - \mathbf{1}_a [\mathbf{1}_a \mathbf{1}_a^\top + s_T^2 I]^{-1} \mathbf{1}_a,$$

where $\mathbf{1}_a = [1, \dots, 1]_{1 \times n_{T,t}^0(a)}^\top$. By the Sherman-Morrison formula, $[\mathbf{1}_a \mathbf{1}_a^\top + s_T^2 I]^{-1} = s_T^{-2} I - \frac{s_T^{-4} \mathbf{1}_a \mathbf{1}_a^\top}{1 + s_T^{-2} n_{T,t}^0(a)}$. Thus the posterior variance is

$$\sigma_{T,t}^2(a) = \frac{1}{1 + s_T^{-2} n_{T,t}^0(a)}. \quad (2.10)$$

Following the arguments in [SKKS10a], we derive the following results. For any $t \leq T$, and an arbitrary sequence $\mathbf{a}_t = \{a_1, a_2, \dots, a_t\}$, we consider fixing this sequence and query the constructed Gaussian processes at these points. Since \mathbf{a}_t is fixed, the entropy $H(\tilde{\mathbf{y}}_t, \mathbf{a}_t) = H(\tilde{\mathbf{y}}_t)$. Since, by definition of a Gaussian process, $\tilde{\mathbf{y}}_t$ follows a multivariate Gaussian distribution,

$$H(\tilde{\mathbf{y}}_t) = \frac{1}{2} \log [(2\pi e)^t \det (K + s_T^2 I)] \quad (2.11)$$

where $K = \left[k_T(a_i, a_j) \right]_{t \times t}$. We can then compute $H(\tilde{\mathbf{y}}_t)$ by

$$\begin{aligned}
H(\tilde{\mathbf{y}}_t) &= H(\tilde{y}_t | \tilde{\mathbf{y}}_{t-1}) + H(\tilde{\mathbf{y}}_{t-1}) \\
&= H(\tilde{y}_t | a_t, \tilde{\mathbf{y}}_{t-1}, \mathbf{a}_{t-1}) + H(\tilde{\mathbf{y}}_{t-1}) \\
&= \frac{1}{2} \log (2\pi e (s_T^2 + \sigma_{T,t-1}^2(a_t))) + H(\tilde{\mathbf{y}}_{t-1}) \\
&= \frac{1}{2} \sum_{\tau=1}^t \log (2\pi e (s_T^2 + \sigma_{T,\tau-1}^2(a_\tau))), \tag{2.12}
\end{aligned}$$

where (2.12) comes from recursively expanding $H(\tilde{\mathbf{y}}_\tau)$. By (2.11) and (2.12),

$$\sum_{\tau=1}^t \log (1 + s^{-2} \sigma_{T,\tau-1}^2(a_\tau)) = \log [\det (s^{-2}K + I)]. \tag{2.13}$$

For the block diagonal matrix K of size $t \times t$, let n_i denote the size of block i and B' ($B' \leq |\mathcal{P}_t|$) be the total number of diagonal blocks up to a time t ($t \leq T$). Then we have

$$\begin{aligned}
\det (s^{-2}K + I) &= \prod_{i=1}^{B'} \det (s^{-2} \mathbf{1} \mathbf{1}^\top + I_{n_i \times n_i}) \\
&= \prod_{i=1}^{B'} (1 + s^{-2} n_i) \leq \left(1 + \frac{s^{-2}t}{B'} \right)^{B'},
\end{aligned}$$

where $\mathbf{1}$ is all-1 vector of proper length. In the above, (1) the equality on the first line uses the determinant of block-diagonal matrix equals to the product of determinant of diagonal blocks, 2) the equality on the last line is due to the matrix determinant lemma, and 3) the inequality on the last line is due to the AM-GM inequality and that $\sum_{i=1}^{B'} n_i = t$.

Next, since $|\mathcal{P}_t| \geq B'$ and $\left(1 + \frac{s^{-2}t}{x} \right)^x$ is increasing with x (on $[1, \infty)$),

$$\det (s^{-2}K + I) \leq \left(1 + \frac{s^{-2}t}{B'} \right)^{B'} \leq \left(1 + \frac{s^{-2}t}{|\mathcal{P}_t|} \right)^{|\mathcal{P}_t|}. \tag{2.14}$$

Therefore, from (2.13) and (2.14),

$$\sum_{\tau=1}^T \log (1 + s^{-2} \sigma_{T,\tau-1}^2(a_\tau)) \leq |\mathcal{P}_T| \log \left(1 + \frac{s^{-2}T}{|\mathcal{P}_T|} \right), \tag{2.15}$$

since arguments after (2.11) hold for all $t \leq T$.

Since the function $h(\lambda) = \frac{\lambda}{\log(1+\lambda)}$ is increasing for non-negative λ , $\lambda \leq \frac{s_T^{-2}}{\log(1+s_T^{-2})} \log(1+\lambda)$ for $\lambda \in [0, s_T^{-2}]$. Since $\sigma_{T,t}(a) \in [0, 1]$ for all a ,

$$\sigma_{T,t}^2(a) \leq \frac{1}{\log(1+s_T^{-2})} \log(1+s_T^{-2}\sigma_{T,t}^2(a)) \quad (2.16)$$

for $t, T = 0, 1, 2, \dots$. Since the partitions are nested, we have that for $T_1 \leq T_2$, $n_{T_1,t}(a) \geq n_{T_2,t}(a)$, and thus $\sigma_{T_1,t}^2(a) \leq \sigma_{T_2,t}^2(a)$. Suppose we query at points a_1, \dots, a_T in the Gaussian process $\mathcal{GP}(0, k_T(\cdot, \cdot))$. Then,

$$\begin{aligned} \sum_{t=1}^T \frac{1}{n_{t-1}(a_t)} &\leq \sum_{t=1}^T \frac{1+s_T^{-2}}{1+s_T^{-2}n_{t-1}(a_t)} \\ &\leq \sum_{t=1}^T \frac{1+s_T^{-2}}{1+s_T^{-2}n_{T,t-1}^0(a_t)} \leq (1+s_T^{-2}) \sum_{t=1}^T \sigma_{T,t-1}^2(a_t) \\ &\leq \frac{1+s_T^{-2}}{\log(1+s_T^{-2})} \sum_{t=1}^T \log(1+s_T^{-2}\sigma_{T,t-1}^2(a_t)) \\ &\leq \frac{1+s_T^{-2}}{\log(1+s_T^{-2})} |\mathcal{P}_T| \log\left(1+s_T^{-2} \frac{T}{|\mathcal{P}_T|}\right), \end{aligned} \quad (2.17)$$

where (2.17) uses (2.10), the second last inequality uses (2.16), and the last inequality uses (2.15). Finally, we optimize over s_T . Since $s_T^{-2} = e - 1$ minimizes $\frac{1+s_T^{-2}}{\log(1+s_T^{-2})}$, we have

$$\sum_{t=1}^T \frac{1}{n_{t-1}(a_t)} \leq e|\mathcal{P}_T| \log\left(1+(e-1)\frac{T}{|\mathcal{P}_T|}\right).$$

The above argument proves (2.6).

Remark 3. *One important insight of our analysis is that this allows us to link the Hoeffding-type concentration term to the posterior variance of the constructed Gaussian processes. This connection is directly shown in (2.10). As we will discuss in Section 2.1.2, we can use this connection to improve the entire learning process via “softening”.*

Next, we sketch the proofs of (2.7) and (2.8).

Proof of (2.7). Consider the partition \mathcal{P}_T at time T . We label the regions of the partitions by $j = 1, 2, \dots, |\mathcal{P}_T|$. Let $t_{j,i}$ be the time when the i -th point in the j -th region in \mathcal{P}_T being selected. Let b_j be the number of points in region j . Since the partitions are nested, we have $1 + n_{t_{j,i}-1}^0(x_{t_{j,i}}) \geq i$ for all i, j . We have, for $T \geq 1$,

$$\sum_{t=1}^T \frac{1}{1+n_{t-1}^0(x_t)} = \sum_{j=1}^{|\mathcal{P}_T|} \sum_{i=1}^{b_j} \frac{1}{1+n_{t_j,i-1}^0(x_{t_j,i}^0)} \leq \sum_{j=1}^{|\mathcal{P}_T|} \sum_{i=1}^{b_j} \frac{1}{i} \quad (2.18)$$

$$\begin{aligned} &\leq \sum_{j=1}^{|\mathcal{P}_T|} (1 + \log b_j) = |\mathcal{P}_T| + \sum_{j=1}^{|\mathcal{P}_T|} \log b_j \\ &= |\mathcal{P}_T| + \log \prod_{j=1}^{|\mathcal{P}_T|} b_j \leq |\mathcal{P}_T| + |\mathcal{P}_T| \log \frac{T}{|\mathcal{P}_T|}, \end{aligned} \quad (2.19)$$

where (2.18) uses $1+n_{t_j,i-1}^0(x_{t_j,i}^0) \geq i$ and (2.19) uses AM-GM inequality and that $\sum_{j=1}^{|\mathcal{P}_T|} b_j = T$.

Proof of (2.8). The idea is similar to that of (2.7). For $0 < \alpha < 1$,

$$\begin{aligned} \sum_{t=1}^T \left(\frac{1}{1+n_{t-1}^0(x_t)} \right)^\alpha &= \sum_{j=1}^{|\mathcal{P}_T|} \sum_{i=1}^{b_j} \left(\frac{1}{1+n_{t_j,i}^0(x_{t_j,i}^0)} \right)^\alpha \\ &\leq \sum_{j=1}^{|\mathcal{P}_T|} \sum_{i=1}^{b_j} \frac{1}{i^\alpha} \leq \sum_{j=1}^{|\mathcal{P}_T|} \frac{1}{1-\alpha} b_j^{1-\alpha} \\ &\leq \frac{1}{1-\alpha} |\mathcal{P}_T|^\alpha T^{1-\alpha}, \end{aligned} \quad (2.20)$$

where (2.20) is due to the Hölder's inequality and that $\sum_{j=1}^{|\mathcal{P}_T|} b_j = T$.

Proof of Theorem 1

Now we are ready to prove Theorem 1. We can split the sum of regrets by

$$\sum_{t=1}^T (f(a^*) - f(a_t)) = \sum_{t=1}^{\lfloor \sqrt{T} \rfloor} (f(a^*) - f(a_t)) + \sum_{t=\lfloor \sqrt{T} \rfloor + 1}^T (f(a^*) - f(a_t)).$$

Also, by Claim 2, with probability at least $1 - \frac{1}{3^{\lfloor \sqrt{T} \rfloor^3}}$, (2.5) holds simultaneously for all $t = \lfloor \sqrt{T} \rfloor + 1, \dots, T$ ($T \geq 2$). Thus for $T \geq 2$, the event

$$\begin{aligned} E_T &= \left\{ \frac{R_T}{T} > \frac{1}{T} \left(\sqrt{T} + \sum_{t=\lfloor \sqrt{T} \rfloor + 1}^T B_t \right) \right\}, \quad \text{where} \\ B_t &:= \left(2L \cdot D(p_{t-1}(a_t)) + 2C \sqrt{\frac{4 \log t}{n_{t-1}(a_t)}} \right) \end{aligned}$$

occurs with probability at most $\frac{1}{3\lfloor\sqrt{T}\rfloor^3}$. Since $\frac{1}{3\lfloor\sqrt{T}\rfloor^3} \sim \frac{1}{3T^{3/2}}$, we know $\sum_{T=2}^{\infty} \mathbb{P}(E_T) < \infty$. By the Borel-Cantelli lemma, we know $\mathbb{P}(\limsup_{T \rightarrow \infty} E_T) = 0$. In other words, with probability 1, E_T occurs finitely many times. Thus, with probability 1, there exists a constant T_0 , such that the event \bar{E}_T (negation of E_T) occurs for all $T > T_0$. Also, from the Cauchy-Schwarz inequality (used below in the second line) and (2.6) (used below in the last line), we know that

$$\begin{aligned} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^T \sqrt{\frac{\log t}{n_{t-1}(a_t)}} &\leq \sum_{t=1}^T \sqrt{\frac{\log t}{n_{t-1}(a_t)}} \leq \sqrt{T \log T} \sqrt{\sum_{t=1}^T \frac{1}{n_{t-1}(a_t)}} \\ &\leq \sqrt{T \log T} \sqrt{e|\mathcal{P}_T| \log \left(1 + (e-1) \frac{T}{|\mathcal{P}_T|}\right)} = \tilde{\mathcal{O}}\left(T^{\frac{1+\gamma}{2}}\right), \end{aligned}$$

where the last equality is from the assumption that $|\mathcal{P}_t| = o(t^\gamma)$ for some $\gamma < 1$. This means

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sqrt{\frac{4 \log t}{n_{t-1}(a_t)}} = 0.$$

In addition, by the assumption that $D(p_t(a)) = o(1)$, we know

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T D(p_{t-1}(a)) = 0.$$

The above two limits give us

$$\lim_{T \rightarrow \infty} \frac{1}{T} \left(\sqrt{T} + \sum_{\lfloor\sqrt{T}\rfloor+1}^T B_t \right) = 0, \quad \text{where} \quad (2.21)$$

$$B_t := \left(2L \cdot D(p_{t-1}(a_t)) + 2C \sqrt{\frac{4 \log t}{n_{t-1}(a_t)}} \right). \quad (2.22)$$

Combining all the facts above, we have $\lim_{T \rightarrow \infty} \frac{R_T}{T} = 0$ with probability 1.

Adaptive partitioning: TUCB shall be implemented using regression trees or incremental regression trees. This naturally leverages the practical advantages of regression trees. Leaves in a regression tree form a partition of the space. Also, a regression tree is designed to fit an underlying function. This leads to an adaptive partitioning where the underlying function values within each region should be relatively similar to each other. We defer the discussion on the implementation we use in our experiments to Section 2.1.3. Please refer to [BFSO84] for more details about regression tree fitting.

The Contextual TreeUCB algorithm

Algorithm 2 Contextual TreeUCB (CTUCB)

- 1: Parameter: $M > 0$, $C > 0$, and tree fitting rule \mathcal{R} .
- 2: **for** $t = 1, 2, \dots, T$ **do**
- 3: Observe context z_t .
- 4: Fit a regression tree f_{t-1} (using rule \mathcal{R}) on observations $\{(z_t, a_t), y_t\}_{t=1}^T$.
- 5: With respect to the partition \mathcal{P}_{t-1} defined by leaves of f_{t-1} , define m_{t-1} and n_{t-1} in (2.1) and (2.2) (over the joint space $\mathcal{Z} \times \mathcal{A}$). Play

$$a_t \in \arg \max_{a \in \mathcal{A}} \{U_t((z_t, a))\},$$

where $U_t(\cdot)$ is defined in (2.3). Ties are broken at random.

- 6: Observe the reward y_t .
-

In this section, we present an extension of Algorithm 1 for the contextual stochastic bandit problem. The contextual stochastic bandit problem is an extension to the stochastic bandit problem. In this problem, at each time, context information is revealed, and the agent chooses an arm based on past experience as well as the contextual information. Formally, the expected payoff function f is defined over the product of the context space \mathcal{Z} and the arm space \mathcal{A} and takes values from $[0, 1]$. Similar to the previous discussions, compactness of the product space and Lipschitzness of the payoff function are assumed. In addition, a mean zero, almost surely bounded noise that is independent of the expected reward function is added to the observed rewards. At each time t , a contextual vector $z_t \in \mathcal{Z}$ is revealed and the agent plays an arm $a_t \in \mathcal{A}$. The performance of the agent following algorithm Alg is measured by the cumulative contextual regret

$$R_T^c(\text{Alg}) = \sum_{t=1}^T f(z_t, a_t^*) - f(z_t, a_t), \quad (2.23)$$

where $f(z_t, a_t^*)$ is the maximal value of f given contextual information z_t . A simple extension of Algorithm 1 can solve the contextual version problem. In particular, in the

contextual case, we partition the joint space $\mathcal{Z} \times \mathcal{A}$ instead of the arm space \mathcal{A} . As an analog to (2.2) and (2.1), we define the corrected count n_t and the corrected average m_t over the joint space $\mathcal{Z} \times \mathcal{A}$ with respect to the partition \mathcal{P}_t of the joint space $\mathcal{Z} \times \mathcal{A}$, and observations in the joint space $((z_1, a_1), y_1, \dots, (z_t, a_t), y_t)$. The guarantee of Algorithm 2 is in Theorem 2.

Theorem 2. *Suppose that the payoff function f defined on a compact doubling metric space $(\mathcal{Z} \times \mathcal{A}, d)$ satisfies $f(z, a) \in [0, 1]$ for all (z, a) and is Lipschitz. If the tree growing rule satisfies requirements 1-4 listed in Theorem 1, then $\lim_{T \rightarrow \infty} \frac{R_T^c(\text{CTUCB})}{T} = 0$ with probability 1.*

Theorem 2 follows from Theorem 1. Since the point scattering inequality holds for any sequence of (context-)arms, we can replace regret with contextual regret and alter Claims 1 and 2 accordingly to prove Theorem 2.

In particular, Claims 1 and 2 extend to the contextual setting, as stated and proved below.

Claim 3. *For any context z , arm a , and time t , with probability at most $\frac{1}{t^4}$, we have:*

$$|m_{t-1}(z, a) - f(z, a)| > L \cdot D(p_{t-1}(z, a)) + C \sqrt{\frac{4 \log t}{n_{t-1}(z, a)}} \quad (2.24)$$

for a constant C .

Proof. First of all, when $t = 1$, this is trivially true by Lipschitzness. Now let us consider the case when $t \geq 2$. Let us use A_1, A_2, \dots, A_t to denote the random variables of arms selected up to time t , Z_1, Z_2, \dots, Z_t to denote the random context up to time t and Y_1, Y_2, \dots, Y_t to denote random variables of rewards received up to time t . Then the random variables $\left\{ \sum_{t=1}^T (f(Z_t, A_t) - Y_t) \right\}$ is a martingale sequence. This is easy to verify since the noise is mean zero and independent. In addition, since there is no randomness in the partition formation (given a sequence of observations), for a fixed a , we have the times $\mathbb{I}[(Z_t, A_i) \in p_{t-1}(z, a)]$ ($i \leq t$) is measurable with respect to $\sigma(Z_1, A_1, Y_1, \dots, Z_t, A_t, Y_t)$. Therefore, the sequence $\left\{ \sum_{i=1}^t (f(Z_i, A_i) - Y_i) \mathbb{I}[(Z_i, A_i) \in p_{t-1}(z, a)] \right\}_{t=1}^T$ is a skipped martingale. Since

skipped martingale is also a martingale, we apply the Azuma-Hoeffding inequality (with sub-Gaussian tails) [Sha11]. For simplicity, we write

$$B_t(z, a) := C \sqrt{\frac{4 \log t}{n_{t-1}(z, a)}} + L \cdot D(p_{t-1}(z, a)), \quad (2.25)$$

$$\mathcal{E}_t^i(z, a) := (Z_i, A_i) \in p_{t-1}(z, a). \quad (2.26)$$

Combining this with Lipschitzness, we get there is a constant C (depends on the a.s. bound of the reward, as a result of Hoeffding inequality), such that

$$\begin{aligned} & \mathbb{P} \{ |m_{t-1}(z, a) - f(z, a)| > B_t(z, a) \} \\ & \leq \mathbb{P} \left\{ \left| \frac{1}{n_{t-1}(z, a)} \sum_{i=1}^{t-1} (f(Z_i, A_i) - Y_i) \mathbb{I}[\mathcal{E}_t^i(z, a)] \right| \right. \\ & \quad \left. + \left| f(z, a) - \frac{1}{n_{t-1}(z, a)} \sum_{i=1}^{t-1} f(Z_i, A_i) \mathbb{I}[\mathcal{E}_t^i(z, a)] \right| \right. \\ & \quad \left. > C \sqrt{\frac{4 \log t}{n_{t-1}(z, a)}} + L \cdot D(p_{t-1}(z, a)) \right\} \leq \frac{1}{t^4}, \end{aligned} \quad (2.27)$$

where (2.27) uses both the Lipschitzness and the Azuma-Hoeffding's inequality. \square

Claim 4. *At any t , with probability at least $1 - \frac{1}{t^4}$, the single step contextual regret satisfies:*

$$f(z_t, a_t^*) - f(z_t, a_t) \leq 2L \cdot D(p_{t-1}(z_t, a_t)) + 2C \sqrt{\frac{4 \log t}{n_{t-1}(z_t, a_t)}}$$

for a constant C . Here a_t^* is the optimal arm for the context z_t .

Proof. By Claim 3, with probability at least $1 - \frac{1}{t^4}$, the following ((2.28) and (2.29)) hold simultaneously,

$$\begin{aligned} & m_{t-1}(z_t, a_t) + C \sqrt{\frac{4 \log t}{n_{t-1}(z_t, a_t)}} + L \cdot D(p_{t-1}(z_t, a_t)) \\ & \geq m_{t-1}(z_t, a_t^*) + \sqrt{\frac{4 \log t}{n_{t-1}(z_t, a_t^*)}} + L \cdot D(p_{t-1}(z_t, a_t^*)) \\ & \geq f(z_t, a_t^*), \end{aligned} \quad (2.28)$$

$$f(z_t, a_t) \geq m_{t-1}(z_t, a_t) - C \sqrt{\frac{4 \log t}{n_{t-1}(z_t, a_t)}} - L \cdot D(p_{t-1}(z_t, a_t)). \quad (2.29)$$

This is true since we first take a one-sided version of Hoeffding-type tail bound in (2.24), and then take a union bound over the two points (z_t, a_t) and (z_t, a_t^*) . This first halves the probability bound and then doubles it. Then we take the complementary event to get (2.28) and (2.29) simultaneously hold with probability at least $1 - \frac{1}{t^4}$. We then take another union bound over time t , as discussed in the main text. Note that throughout the proof, we do not need to take union bounds over all arms or all regions in the partition.

Equation 2.28 holds by algorithm definition. Otherwise we will not select a_t at time t . Combine (2.28) and (2.29), and we get

$$\begin{aligned} & f(z_t, a_t^*) - f(z_t, a_t) \\ &= f(z_t, a_t^*) - m_{t-1}(z_t, a_t) + m_{t-1}(z_t, a_t) - f(z_t, a_t) \\ &\leq 2C \sqrt{\frac{4 \log t}{n_{t-1}(z_t, a_t)}} + 2L \cdot D(p_{t-1}(z_t, a_t)). \end{aligned}$$

□

Use Cases of Point Scattering Inequalities

Recover Previous Bounds

In this section, we give examples of using the point scattering inequalities to derive regret bounds for other algorithms. For our purpose of illustrating the point scattering inequalities, the discussed algorithms are simplified. We also assume that the reward and the sub-Gaussianity are properly scaled so that the parameter before the Hoeffding-type concentration term is 1.

The UCB1 algorithm The classic UCB1 algorithm [ACBF02] assumes a finite set of arms, each having a different reward distribution. Following our notation, at time t , the UCB1 algorithm plays

$$a_t \in \arg \max_a \left\{ m_{t-1}(a) + \sqrt{\frac{2 \log T}{n_{t-1}(a)}} \right\}. \quad (2.30)$$

(a) CNN architecture for SVHN. A value with * means that this parameter is tuned, and the batch-normalization layer uses all Tensorflow’s default settings.

Layer	Hyperparameters	values
Conv1	conv1-kernel-size	*
	conv1-number-of-channels	200
	conv1-stride-size	(1,1)
MaxPooling1	pooling1-size	(3,3)
	pooling1-stride	(1,1)
Conv2	conv2-kernel-size	*
	conv2-number-of-channels	200
	conv2-stride-size	(1,1)
MaxPooling2	pooling2-size	(3,3)
	pooling2-stride	(2,2)
Conv3	conv3-kernel-size	(3,3)
	conv3-number-of-channels	200
	conv3-stride-size	(1,1)
AvgPooling3	pooling3-size	(3,3)
	pooling3-stride	(1,1)
Dense	batch-normalization	default
	number-of-hidden-units	512
	dropout-rate	0.5

(b) Hyperparameter search space. β_1 and β_2 are parameters for the AdamOptimizer [KB15]. The learning rate is discretized in the following way: from 1e-6 to 1 (including the end points), we log-space the learning rate into 50 points, and from 1.08 to 5 (including the end points) we linear-space the learning rate into 49 points.

Hyperparameters	Range
conv1-kernel-size	{1, 2, ..., 7}
conv2-kernel-size	{1, 2, ..., 7}
β_1 & β_2	{0, 0.05, ..., 1}
learning-rate	1e-6 to 5
training-iteration	{300, 400, ..., 1500}

Table 2.1: Settings for the SVHN experiments.

(a) CNN architecture for CIFAR-10. A value with * means that this parameter is tuned, and the batch-normalization layer uses all Tensorflow’s default setting.

Layer	Hyperparameters	values
Conv1	conv1-kernel-size	*
	conv1-no.-of-channels	200
	conv1-stride-size	(1,1)
MaxPooling1	pooling1-size	*
	pooling1-stride	(1,1)
Conv2	conv2-kernel-size	*
	conv2-no.-of-channels	200
	conv2-stride-size	(1,1)
MaxPooling2	pooling2-size	*
	pooling2-stride	(2,2)
Conv3	conv3-kernel-size	*
	conv3-no.-of-channels	200
	conv3-stride-size	(1,1)
AvgPooling3	pooling3-size	*
	pooling3-stride	(1,1)
	pooling3-padding	“same”
Dense	batch-normalization	default
	no.-of-hidden-units	512
	dropout-rate	0.5

(b) Hyperparameter search space. β_1 and β_2 are parameters for the Adamoptimizer. The learning rate is discretized in the following way: from 1e-6 to 1 (including the end points), we log-space the learning rate into 50 points, and from 1.08 to 5 (including the end points) we linear-space the learning rate into 49 points. The learning-rate-reduction parameter is how many times the learning rate is going to be reduced by a factor of 10. For example, if the total training iteration is 200, the learning-rate is 1e-6, and the learning-rate-reduction is 1, then for the first 100 iteration the learning rate is 1e-6, and the for last 100 iterations the learning rate is 1e-7.

Hyperparameters	Range
conv1-kernel-size	{1, 2, ..., 7}
conv2-kernel-size	{1, 2, ..., 7}
conv3-kernel-size	{1, 2, 3}
pooling1-size & pooling2-size	{1, 2, 3}
pooling3-size	{1, 2, ..., 6}
β_1 & β_2	{0, 0.05, ..., 1}
learning-rate	1e-6 to 5
learning-rate-redeuction	{1,2,3}
training-iteration	{200, 400, ..., 3000}

Table 2.2: Settings for CIFAR-10 experiments.

Indeed, this equation can be interpreted as (2.4) under the discrete 0-1 metric: two points are distance zero if they coincide and distance 1 otherwise. Then from the point scattering inequality (2.6), we get for UCB1

$$\begin{aligned}\mathbb{E}[R_T(\text{UCB1})] &= \mathcal{O}\left(\sum_{t=1}^T \sqrt{\frac{\log T}{n_{t-1}(a_t)}}\right) \\ &= \mathcal{O}\left(\sqrt{T \log T} \sqrt{\sum_{t=1}^T \frac{1}{n_{t-1}(a_t)}}\right) = \tilde{\mathcal{O}}\left(\sqrt{K \cdot T}\right),\end{aligned}$$

where K is number of arms in the problem. This matches the gap-independent (independent of the reward gap between an arm and the optimal arm) bound derived using traditional methods in UCB1 algorithm [ACBF02, BCB⁺12]. In this analysis, we apply the point scattering inequality with the partition \mathcal{P}_t being the set of arms at all t .

Finite Time Bound for Lipschitz Bandits and Lipschitz RL. As shown in Claim 2, the single step regret is bounded by a Hoeffding-type concentration and the diameter of selected region (due to Lipschitzness). Since the point scattering inequalities provide a bound of the overall summation of the Hoeffding terms, we can design and analyze many partition-based Lipschitz algorithms using point scattering inequalities. We can do this since the partitioning is up to our choice. Examples include the **UniformMesh** algorithm discussed by [KSU08], and partition-based Lipschitz reinforcement learning algorithm recently studied (e.g., [NYW19]).

Hierarchical Bayesian Method for Lipschitz Bandits

Existing Lipschitz bandit algorithms (e.g., [KSU08]) partition the arm space into disjoint bins. Based on this partition, arms in two different bin do not give information about each other, and all arms within the same bins are viewed as the same. This implicit assumption, however, is obviously untrue. On the other hand, imposing a strong prior on the reward function would break the Lipschitzness assumption. To simultaneously address the above two difficulties, we link the learned tree (or partition) to a Bayesian model in light of our analysis of (2.6). This new viewpoint allows us to “soften” the entire model using a hierarchical Bayesian method.

Formally, at each time t , we consider the following hierarchical Bayesian problem with respect to the learned partition \mathcal{P}_t . Note that this hierarchical Bayesian model is updated whenever we update the partition. This is roughly the same as make a finite partition and treat each bin as an arm, and do not impose extra structures on the reward function. Let \mathcal{P}_t be the learnt partition such that each bin is a rectangle. Then the kernel function is defined as

$$\tilde{k}_T(\cdot, \cdot) = \sum_{p \in \mathcal{P}_T} \tilde{k}_T^{(p)}(\cdot, \cdot), \quad (2.31)$$

where p are regions in \mathcal{P}_T , and $\tilde{k}_T^{(p)}(\cdot, \cdot)$ is defined as follows. For a partition $p = \prod_{i=1}^d [a_i, b_i]$, define

$$\tilde{k}_T^{(p)}(\cdot, \cdot) = \prod_{i=1}^d \tilde{k}_T^{(p,i)}(\cdot, \cdot), \quad \text{where} \quad (2.32)$$

$$\tilde{k}_T^{(p,i)}(\mathbf{x}, \mathbf{x}') = \left[1 + \exp \left(-\alpha_T \left(\Delta_i - \frac{b_i - a_i}{2} \right) \right) \right]^{-1}, \quad (2.33)$$

$$\Delta_i = \max \left\{ \left| \mathbf{x}_i - \frac{a_i + b_i}{2} \right|, \left| \mathbf{x}'_i - \frac{a_i + b_i}{2} \right| \right\}, \quad (2.34)$$

where \mathbf{x}_i (resp. \mathbf{x}'_i) are the i -th entry of \mathbf{x}_i (resp. \mathbf{x}'), and $\alpha_T > 0$ are parameters that controls how smooth are the smoothed tree metrics. Given a learned partition $\mathcal{P}_T = \{p_1, p_2, \dots, p_K\}$, where $p_j = \prod_{i=1}^d [a_j^{(i)}, b_j^{(i)}]$, we construct the following hierarchical Bayesian model

$$\tilde{a}_j^{(i)} \sim \mathcal{N}(a_j^{(i)}, \sigma^2), \text{ for all } i, j; \quad \tilde{b}_j^{(i)} \sim \mathcal{N}(b_j^{(i)}, \sigma^2), \text{ for all } i, j \quad (2.35)$$

$$\tilde{k}_T = \sum_{j=1}^K \tilde{k}_T^{(p_j)}, \text{ where } \tilde{k}_T^{(p_j)} \text{ is defined respect to } \prod_{i=1}^d [\tilde{a}_j^{(i)}, \tilde{b}_j^{(i)}]$$

$$f \sim \mathcal{GP} \left(0, \tilde{k}_T(\cdot, \cdot) \right) \quad (2.36)$$

$$y = f + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma_y^2). \quad (2.37)$$

This hierarchical model has several advantages: (1) It respect Lipschitzness. As we collect more observations, the partition can grow arbitrarily fine, and the approximation

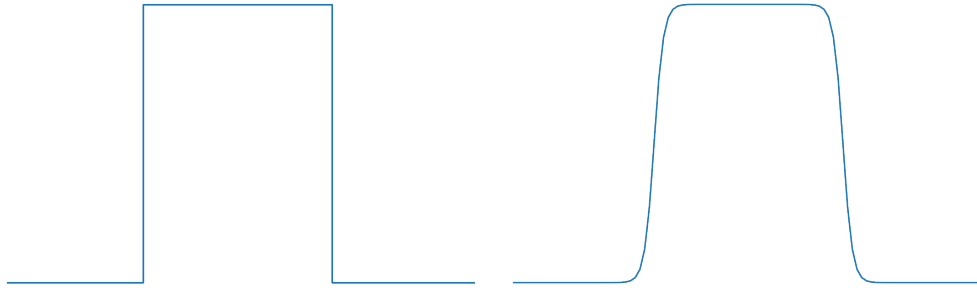


Figure 2.2: The left subfigure is the metric learned by Algorithm 1 (2.9). The right subfigure is the smoothed version of this learned metric.

can be arbitrarily close to an extract indicator function. Because of this, the no prior smoothness assumption on the true (unknown) reward function is needed. (2) It treats arms within the same bin differently, and can use information across bins.

Going back to bandit learning process, we can replace the mean and/or confidence intervals of UCB index with the posteriors of this hierarchical bayesian model. As we discussed in Remark 3, a key insight of our analysis is the link between the Hoeffding-type concentration interval to the posterior variance of the Gaussian processes, which allows us to do this principled substitute. In Section 2.1.3, we empirically study this hierarchical Bayesian model.

2.1.3 Empirical Study

Since the TreeUCB algorithm imposes only mild constraints on tree formation, we use greedy decision tree splitting to fit the reward function, using the following splitting rule: we find the split that maximizes the reduction in the Mean Absolute Error (MAE), and we stop growing the tree once the maximal possible reduction is below 0.001.

Gaussian Processes with Learned Kernel

In this section, we compare several baselines, including piecewise constant estimates (within each bin), a Gaussian process regression with box kernel (left subfigure in Figure 2.2) and

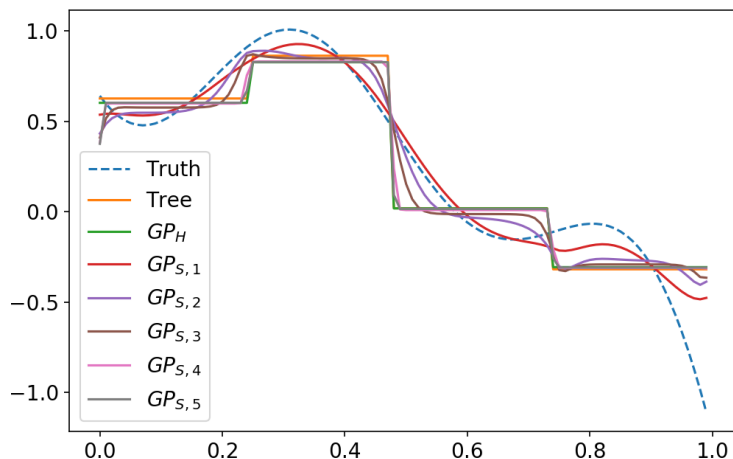


Figure 2.3: The estimates for a function with respect to a given partition. The “Tree” line is directly averaging within each partition. The “ GP_H ” line is the learned posterior GP mean function using the “hard metric.” The lines “ $GP_{S,1} - GP_{S,5}$ ” are 5 learned posterior GP mean functions using the “soft metric” (Eq. (2.32) - (2.34)).

Gaussian process regression with softened box kernel (right subfigure in Figure 2.2). The splitting procedure is the same for all methods, so the partitions are the same for the methods. Our results, shown in Figure 2.3, demonstrates a transition from the hardness of the piecewise constant estimate to the softness of the Gaussian process regression with the softened kernel. This justifies the “softening” discussed in Section 2.1.2. The Gaussian process kernel parameters for $GP_{S,1}, GP_{S,2}, GP_{S,3}, GP_{S,4}, GP_{S,5}$, namely α_T in Eq. (2.33), were set to 10, 50, 100, 500, 1000 respectively.

Application to Neural Network Tuning

One application of stochastic bandit algorithms is zeroth order optimization. In this section, we apply TUCB to tuning neural networks. In this setting, we treat the hyperparameter configurations (e.g., learning rate, network architecture) as the arms of the bandit, and use validation accuracy as reward. The task is to select a hyperparameter configuration and train the network to observe the validation accuracies, and find the best hyperparameter configuration rapidly. This experiment shows that TUCB can compete with the state-of-the-art tuning methods on such hard real-world tasks.

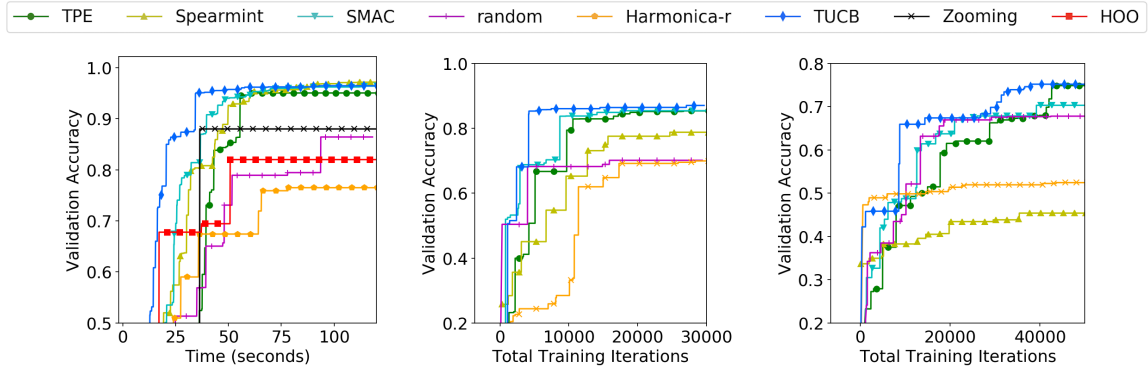


Figure 2.4: For MNIST, each plot is averaged over 10 runs. For SVHN and CIFAR-10, each plot is averaged over 5 runs. The implementation of TUCB here uses the scikit-learn package [PVG⁺11a]. In the left-most subplot, x-axis is time (in seconds). This shows TUCB’s scalability, since TUCB’s curve goes up the fastest. In the leftmost sub-figure, we use clock time as cost measure.

The architecture and the hyperparameter space for the simple Multi-Layer Perceptron (MLP) for the MNIST dataset are: in the feed-forward direction, there are the input layer, the fully connected hidden layer with dropout ensemble, and then the output layer. The hyperparameter search space is five dimensional, including number of hidden neurons (range [10, 784]), learning rate ([0.0001, 4]), dropout rate ([0.1, 0.9]), batch size ([10, 500]), and number of iterations ([30, 243]).

The details of the CNN setting for SVHN and CIFAR-10 can be found in Tables 2.1 and 2.2. The results are found in Figure 2.4, indicating that TUCB outperforms existing state-of-the-art software packages for tuning neural network methods.

2.1.4 Conclusion

We propose the TreeUCB and the Contextual TreeUCB frameworks that use decision trees (regression trees) to flexibly partition the arm space and the context-arm space as an Upper Confidence Bound strategy is played across the partition regions. We also provide regret analysis via the point scattering inequalities. We provide implementations using decision trees that learn the partition. TUCB is competitive with the state-of-the-art hyperparameter optimization methods in hard tasks like neural-net tuning, and could save substantial

computing resources. This suggests that, in addition to random search and Bayesian optimization methods, more bandit algorithms should be considered as benchmarks for difficult real-world problems such as neural network tuning.

2.2 Bandits for BMO Functions

2.2.1 Introduction

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.

Related Works

Bandit problems in different settings have been actively studied since as far back as Thompson [Tho33]. Upper confidence bound (UCB) algorithms remain popular [Rob52, LR85, Aue02] among the many approaches for (stochastic) bandit problems [SKKS10a, AYPS11, AG12, BS12, SS14]. Various extensions of upper confidence bound algorithms have been studied. Some works use KL-divergence to construct the confidence bound [LR85, GC11, MMS11], and some works include variance estimates within the confidence bound [AMS09, AO10]. UCB is also used in the contextual setting [LCLS10, KO11, Sli14].

Perhaps Lipschitz bandits are closest to BMO bandits. The Lipschitz bandit problem was termed “continuum-armed bandits” in early stages [Agr95a]. 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 [Kle05] proves a $\Omega(T^{2/3})$ lower bound and proposes a $\tilde{\mathcal{O}}(T^{2/3})$ algorithm. Under other extra

conditions on top of Lipschitzness, regret rate of $\tilde{O}(T^{1/2})$ was achieved [Cop09, AOS07]. For general (doubling) metric spaces, the Zooming bandit algorithm [KSU08] and Hierarchical Optimistic Optimization algorithm [BMSS11] 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. [BSY11] studied Lipschitz bandits for differentiable rewards, which enables algorithms to run without explicitly knowing the Lipschitz constants. The idea of robust mean estimators [BCBL13, B⁺65, AMS99] was applied to the Lipschitz bandit problem to cope with heavy-tail rewards, leading to the development of a near-optimal algorithm [LWHZ19]. Lipschitz bandits with an unknown metric, where a clustering is used to infer the underlying unknown metric, has been studied by Wanigasekara and Yu [WY19a]. Lipschitz bandits with discontinuous but bounded rewards were studied by Krishnamurthy et al. [KLSZ19].

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 [KLSZ19], which may allow discontinuities but require boundedness. To the best of our knowledge, we are the first to study the bandit learning problem for BMO functions.

2.2.2 Preliminaries

We review the concept of (rectangular) Bounded Mean Oscillation (BMO) in Euclidean space [Fef79, SM93].

Definition 3. (*BMO Functions*) Let (\mathbb{R}^d, μ) be the Euclidean space with the Lebesgue measure. Let $L_{loc}^1(\mathbb{R}^d, \mu)$ denote the space of measurable functions (on \mathbb{R}^d) that are locally integrable with respect to μ . A function $f \in L_{loc}^1(\mathbb{R}^d, \mu)$ 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 $Q \subset \mathbb{R}^d$,

$$\frac{1}{\mu(Q)} \int_Q |f - \langle f \rangle_Q| d\mu \leq C_f, \quad \langle f \rangle_Q := \frac{\int_Q f d\mu}{\mu(Q)}. \quad (2.38)$$

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 part.

A BMO function can be discontinuous and unbounded. The function in Figure 2.5 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 3), which plays a central role in our paper.

Theorem 3 (John-Nirenberg). *Let μ be the Lebesgue measure. Let $f \in BMO(\mathbb{R}^d, \mu)$. Then there exists constants C_1 and C_2 , such that, for any hypercube $q \subset \mathbb{R}^d$ and any $\lambda > 0$,*

$$\mu\left(\left\{x \in q : \left|f(x) - \langle f \rangle_q\right| > \lambda\right\}\right) \leq C_1 \mu(q) \exp\left\{\frac{-\lambda}{C_2 \|f\|}\right\}. \quad (2.39)$$

The John-Nirenberg inequality dates back to at least John [Joh61], and a proof is provided in Appendix A.8.

As shown in Appendix A.8, $C_1 = e$ 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 [Ler13, SV17], and it is also conjectured that C_2 and C_1 might be independent of dimension [CSS12]. 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.

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 4.

Theorem 4. [Ass83]. *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.

2.2.3 Problem Setting: BMO Bandits

The goal of a stochastic bandit algorithm is to exploit the current information, and explore the space efficiently. 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:

$$f : [0, 1]^d \rightarrow \mathbb{R} \quad \text{where} \quad f \in BMO([0, 1]^d, \mu). \quad (2.40)$$

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 , $|\mathcal{E}_a| \leq D_{\mathcal{E}}$ almost surely for some constant $D_{\mathcal{E}}$ (**N1**). Our results generalize to the setting with sub-Gaussian noise [Sha11]. 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 $(a_1, y_1, \dots, a_{t-1}, y_{t-1})$, 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, \dots, a_T , we use δ -regret incurred up to time T as the performance measure (Definition 4).

Definition 4. (*δ -regret*) Let $f \in BMO([0, 1]^d, \mu)$. A number $\delta \geq 0$ is called *f -admissible* if there exists a real number z_0 that satisfies

$$\mu(\{a \in [0, 1]^d : f(a) > z_0\}) = \delta. \quad (2.41)$$

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

$$F^\delta := \left\{ z \in \mathbb{R} : \mu(\{a \in [0, 1]^d : f(a) > z\}) = \delta \right\}. \quad (2.42)$$

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

$$r_t^\delta := \max\{0, f^\delta - \mathbb{E}_t[f(A_t)]\}, \quad (2.43)$$

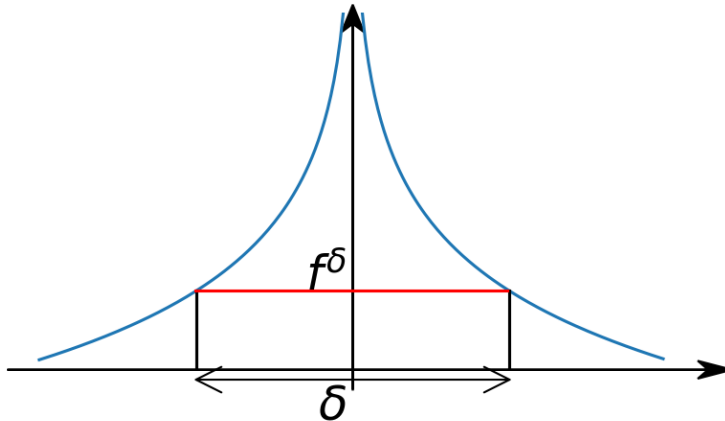


Figure 2.5: Graph of $f(x) = -\log(|x|)$, with δ and f^δ annotated. This function is an unbounded BMO function.

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 2.5 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**.

Remark 4. *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 5 (Connection to bandits with heavy-tails). *In the definition of bandits with heavy tails [BCBL13, MY16, SYKL18, LWHZ19], 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 2.5 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 2. *Let f be the reward function. For any f -admissible $\delta \geq 0$, 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 $\langle f \rangle_{[0, 1]^d} = 0$.

2.2.4 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:

$$Q_{\mathbb{R}^d} := \left\{ \prod_{i=1}^d [m_i 2^{-k}, m_i 2^{-k} + 2^{-k}) \right\} \quad (2.44)$$

where Π is the Cartesian product, and $m_1, \dots, m_d, k \in \mathbb{Z}$. Dyadic cubes of $[0, 1]^d$ is $Q_{[0, 1]^d} := \{q \in Q_{\mathbb{R}^d} : q \subset [0, 1]^d\}$. As a concrete example, dyadic cubes of $[0, 1]^2$ are

$$\{[0, 1)^2, [0, 0.5)^2, [0.5, 1)^2, [0.5, 1) \times [0, 0.5), \dots\}.$$

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 $(a_1, y_1, a_2, y_2, \dots, a_t, y_t)$ be the observations received up to time t . We define

- the cube count $n_t : \mathcal{Q}_t \rightarrow \mathbb{R}$, such that for $q \in \mathcal{Q}_t$

$$n_t(q) := \sum_{i=1}^{t-1} \mathbb{I}_{[a_i \in q]}; \quad \tilde{n}_t(q) := \max(1, n_t(q)). \quad (2.45)$$

- the cube average $m_t : \mathcal{Q}_t \rightarrow \mathbb{R}$, such that for $q \in \mathcal{Q}_t$

$$m_t(q) := \begin{cases} \frac{\sum_{i=1}^{t-1} y_i \mathbb{I}_{[a_i \in q]}}{n_t(q)}, & \text{if } n_t(q) > 0; \\ 0, & \text{otherwise.} \end{cases} \quad (2.46)$$

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

$$Q_t \in \arg \max_{q \in \mathcal{Q}_t} U_t(q), \quad \text{where} \quad (2.47)$$

$$U_t(q) := m_t(q) + H_t(q) + J(q), \quad (2.48)$$

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

$$J(q) := \lceil \log(\mu(q)/\eta) \rceil_+,$$

$$\Psi := \max \{ \log(T^2/\epsilon), 2 \log_2(1/\eta) \}, \quad (2.49)$$

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 5 in Appendix A.2.1). 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 $q \in \mathcal{Q}_t$

$$H_t(q) \geq J(q), \quad \text{or equivalently} \\ \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\tilde{n}_t(q)}} \geq [\log(\mu(q)/\eta)]_+. \quad (2.50)$$

In (2.50), 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 (2.50) 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:

$$Q_t \in \arg \max_{q \in \mathcal{Q}_{t-1}} U_t(q). \quad (2.51)$$

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 3. Analysis of Algorithm 3 is found in Section 2.2.4, which also provides some tools for handling δ -regret. Then in Section 2.2.5, we provide an improved algorithm that exhibits a stronger performance guarantee.

Algorithm 3 Bandit-BMO-Partition (Bandit-BMO-P)

- 1: 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./
- 2: Algorithm parameters: $\eta > 0$, $\epsilon > 0$, T .
/** T is the time horizon. ϵ and η are parameters./
- 3: **for** $t = 1, 2, \dots, T$ **do**
- 4: Let m_t and n_t be defined as in (2.46) and (2.45).
- 5: 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 (2.48).

- 6: Play arm $A_t \in Q_t$ uniformly at random. Observe Y_t .
 - 7: Update the partition \mathcal{Q}_t to \mathcal{Q}_{t+1} according to (2.50).
-

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 5. *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 3 up to time T satisfies*

$$\sum_{t=1}^T r_t^\delta \lesssim_d \tilde{\mathcal{O}}\left(\sqrt{T|\mathcal{Q}_T|}\right), \quad (2.52)$$

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

From uniform tie-breaking, we have

$$\mathbb{E}[f(A_t)|\mathcal{F}_t] = \frac{1}{\mu(Q_t)} \int_{a \in Q_t} f(a) da = \langle f \rangle_{Q_t}, \quad (2.53)$$

$$\mathcal{F}_t = \sigma(Q_1, A_1, Y_1, \dots, Q_{t-1}, A_{t-1}, Y_{t-1}, Q_t), \quad (2.54)$$

where \mathcal{F}_t is the σ -algebra generated by random variables $Q_1, A_1, Y_1, \dots, Q_{t-1}, A_{t-1}, Y_{t-1}, Q_t$ – 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 $t \leq T$, and $q \in Q_t$, we define

$$\mathcal{E}_t(q) := \left\{ \left| \langle f \rangle_q - m_t(q) \right| \leq H_t(q) \right\}, \quad (2.55)$$

$$H_t(q) = \frac{(\Psi + D_{\mathcal{E}}) \sqrt{2 \log(2T^2/\epsilon)}}{\sqrt{\tilde{n}_t(q)}}. \quad (2.56)$$

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 3 shows that $\mathcal{E}_t(q)$ happens with high probability for any t and q .

Lemma 3. *With probability at least $1 - \frac{\epsilon}{T}$, the event $\mathcal{E}_t(q)$ holds for any $q \in Q_t$ at any time t .*

To prove Lemma 3, we apply a variation of Azuma’s inequality [Vu02, TV15]. We also need some additional effort to handle the case when a cube q contains no observations. The details are in Appendix A.2.

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

Lemma 4. *Recall $J(q) = \log(\mu(q)/\eta)$. For any partition \mathcal{Q} of $[0, 1]^d$, there exists $q \in \mathcal{Q}$, such that*

$$f^\delta - \langle f \rangle_q \leq J(q), \quad (2.57)$$

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

In the proof of Lemma 4, 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 4, there exists a “good” cube \tilde{q}_t (at any time $t \leq T$), such that (2.57) is true for \tilde{q}_t . Let δ be an arbitrary number satisfying (1) $\delta > |\mathcal{Q}_T|\eta$ and (2) δ is f -admissible. Then under event $\mathcal{E}(\tilde{q}_t)$,

$$\begin{aligned} f^\delta &= \left(f^\delta - \langle f \rangle_{\tilde{q}_t} \right) + \left(\langle f \rangle_{\tilde{q}_t} - m_t(\tilde{q}_t) \right) + m_t(\tilde{q}_t) \\ &\stackrel{\textcircled{1}}{\leq} J(\tilde{q}_t) + H_t(\tilde{q}_t) + m_t(\tilde{q}_t), \end{aligned} \quad (2.58)$$

where $\textcircled{1}$ uses Lemma 4 for the first brackets and Lemma 3 (with event $\mathcal{E}_t(\tilde{q}_t)$) for the second brackets.

The event where all “good” cubes and all cubes we select (for $t \leq T$) have nice estimates, namely $\left(\bigcap_{t=1}^T \mathcal{E}_t(\tilde{q}_t) \right) \cap \left(\bigcap_{t=1}^T \mathcal{E}_t(Q_t) \right)$, occurs with probability at least $1 - 2\epsilon$. This result comes from Lemma 3 and a union bound, and we note that $\mathcal{E}_t(q)$ depends on ϵ (and T), as in (2.56). Under this event, from (2.55) we have $\left| \langle f \rangle_{Q_t} - m_t(Q_t) \right| \leq H_t(Q_t)$. This and (2.53) give us

$$\mathbb{E}[f(A_t)|\mathcal{F}_t] = \langle f \rangle_{Q_t} \geq m_t(Q_t) - H_t(Q_t). \quad (2.59)$$

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

$$\begin{aligned} & f^\delta - \mathbb{E}[f(A_t)|\mathcal{F}_t] \\ & \stackrel{\textcircled{1}}{\leq} m_t(\tilde{q}_t) + H_t(\tilde{q}_t) + J(\tilde{q}_t) - m_t(Q_t) + H_t(Q_t) \\ & \stackrel{\textcircled{2}}{\leq} m_t(Q_t) + H_t(Q_t) + J(Q_t) - m_t(Q_t) + H_t(Q_t) \\ & = 2H_t(Q_t) + J(Q_t) \leq 3H_t(Q_t), \end{aligned} \quad (2.60)$$

where $\textcircled{1}$ uses (2.58) for the first three terms and (2.59) for the last three terms, $\textcircled{2}$ uses that $U_t(Q_t) \geq U_t(\tilde{q}_t)$ since Q_t maximizes the index $U_t(\cdot)$ according to (2.51), and the last inequality uses the rule (2.50).

Next, we use Lemma 1 to link the number of cubes up to a time t to the Hoeffding-type tail bound in (2.60). Intuitively, this bound (Lemma 1) states that the numbers of points within the cubes grows fast enough to be bounded by a function of the number of cubes.

We can apply Lemma 1 and the Cauchy-Schwarz inequality to (2.60) to prove Theorem 5. The details can be found in Appendix A.4.

2.2.5 Achieve Poly-log Regret via Zooming

In this section we study an improved version of the previous section that uses the Zooming machinery [KSU08, Sli14, BMSS11]. Similar to Algorithm 3, 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 (2.48):

$$U_t(q) := m_t(q) + H_t(q) + J(q) \tag{2.61}$$

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 2.6 gives examples of terminal cubes, pre-parent cubes and parent cubes.

Algorithm Description

Pick **zooming rate** $\alpha \in \left(0, \frac{(\Psi + D_E)\sqrt{2\log(2T^2/\epsilon)}}{\log(M_d/\eta)}\right]$. 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

$$\begin{cases} \frac{(\Psi + D_E)\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{n_{warm}}} \geq \alpha \log\left(\frac{M_d}{\eta}\right) \\ \frac{(\Psi + D_E)\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{n_{warm}+1}} < \alpha \log\left(\frac{M_d}{\eta}\right) \end{cases}. \quad (2.62)$$

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

$$\frac{(\Psi + D_E)\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\tilde{n}_t(Q^{ter})}} \geq \alpha \log\left(\frac{M_d\mu(Q^{ter})}{\eta}\right) \quad (2.63)$$

for any terminal cube Q^{ter} . If (2.63) 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 (2.63). We choose α to be smaller than $\frac{(\Psi + D_E)\sqrt{2\log(2T^2/\epsilon)}}{\log(M_d/\eta)}$ so that (2.63) can be satisfied with $\tilde{n}_t(Q^{ter}) = 1$ 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:

$$\frac{(\Psi + D_E)\sqrt{2\log(2T^2/\epsilon)}}{\sqrt{\tilde{n}_t(Q^{par})}} < \alpha \log\left(\frac{M_d\mu(Q^{par})}{\eta}\right). \quad (2.64)$$

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 4.

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 4 achieves $\tilde{O}(\text{poly-log}(T))$ δ -regret with high probability (formally stated in Theorem 6).

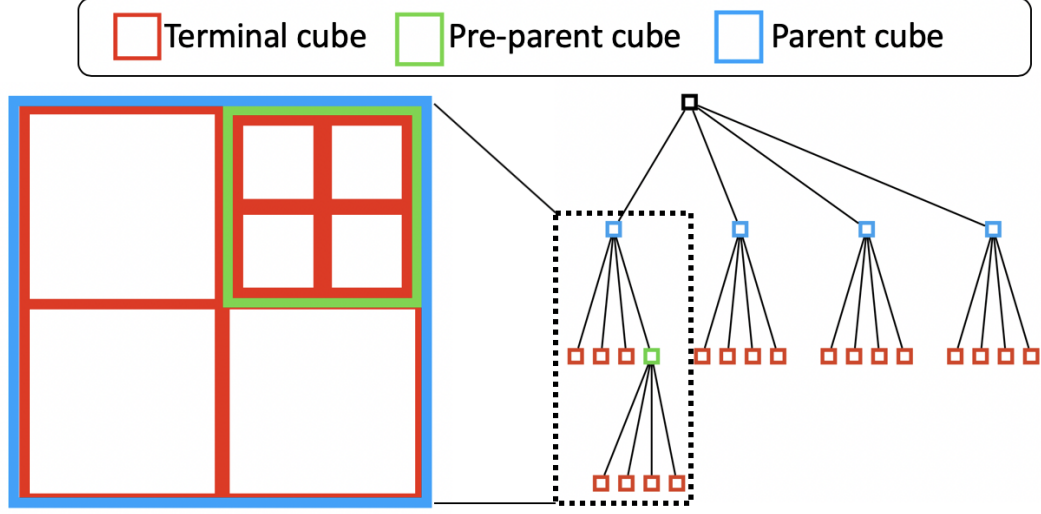


Figure 2.6: Example of terminal cubes, pre-parent and parent cubes.

Algorithm 4 Bandit-BMO-Zooming (Bandit-BMO-Z)

1: Problem intrinsics: $\mu(\cdot)$, $D_{\mathcal{E}}$, d , M_d .

/** $\mu(\cdot)$, $D_{\mathcal{E}}$, d , M_d are same as those in Algorithm 3./

2: Algorithm parameters: $\eta, \epsilon, T > 0$, and $\alpha \in \left(0, \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\log(M_d/\eta)}\right]$.

/** η, ϵ, T are same as those in Algorithm 3. α is the zooming rate./

3: Initialize: let $\mathcal{Q}_0 = [0, 1)^d$. Play warm-up phase (2.62).

4: **for** episode $t = 1, 2, \dots, T$ **do**

5: Let m_t, n_t, U_t be defined as in (2.46), (2.45) and (2.61).

6: Select parent cube $Q_t \in \mathcal{Q}_t$ such that:

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

7: **for** $j = 1, 2, \dots, M_d$ **do**

8: Locate the j -th direct sub-cube of Q_t : Q_j^{sub} .

9: Play $A_{t,j} \in Q_j^{sub}$ uniformly at random, and observe $Y_{t,j}$.

10: Update the collection of dyadic cubes \mathcal{Q}_t to \mathcal{Q}_{t+1} according to (2.63).

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

$$\sum_{i=1}^{M_d} \mathbb{E}_t[f(A_{t,i})] = M_d \langle f \rangle_{Q_t}, \quad (2.65)$$

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

$$\sum_{i=1}^{M_d} \Delta_{t,i}^\delta = M_d (f^\delta - \langle f \rangle_{Q_t}). \quad (2.66)$$

The quantity $\sum_{i=1}^{M_d} \Delta_{t,i}^\delta$ is the δ -regret incurred during episode t . We will bound (2.66) using tools in Section 2.2.4. In order to apply Lemma 4, we need to show that the parent cubes form of partition of the arm space (Proposition 2).

Proposition 2. *At any episode t , the collection of parent cubes forms a partition of the arm space.*

Since the parent cubes in Q_t form a partition of the arm space, we can apply Lemma 4 to get the following. For any episode t , there exists a parent cube q_t^{\max} , such that

$$f^\delta \leq \langle f \rangle_{q_t^{\max}} + \log(\mu(q_t^{\max})/\eta). \quad (2.67)$$

Let us define $\tilde{\mathcal{E}}_T := \left(\bigcap_{t=1}^T \mathcal{E}_t(q_t^{\max}) \right) \cap \left(\bigcap_{t=1}^T \mathcal{E}_t(Q_t) \right)$, where $\mathcal{E}_t(q_t^{\max})$ and $\mathcal{E}_t(Q_t)$ are defined in (2.55). By Lemma 3 and another union bound, we know the event $\tilde{\mathcal{E}}_T$ happens with probability at least $1 - 2\epsilon$.

Since each episode creates at most a constant number of new cubes, we have $|Q_t| = \mathcal{O}(t)$. Using the argument we used for (2.60), we have that at any $t \leq T$, for any $\delta > \eta|Q_t|$ that is f -admissible, under event $\tilde{\mathcal{E}}_T$,

$$\begin{aligned} \sum_{i=1}^{M_d} \Delta_{t,i}^\delta &= M_d \left(f^\delta - \langle f \rangle_{Q_t} \right) \\ &\leq M_d \left(2 \frac{(\Psi + D_{\mathcal{E}}) \sqrt{2 \log(2T^2/\epsilon)}}{\sqrt{\tilde{n}_t(Q_t)}} + \log \left(\frac{M_d \mu(Q_t)}{\eta} \right) \right) \\ &\leq M_d (1 + 2\alpha) \log \left(\frac{M_d \mu(Q_t)}{\eta} \right), \end{aligned} \quad (2.68) \quad (2.69)$$

where (2.68) uses (2.66) and the last inequality uses (2.64).

Next, we extend some definitions from Zooming bandit for Lipschitz functions [KSU08], to handle the δ -regret setting. Firstly, we define the set of (λ, δ) -optimal arms as

$$\mathcal{X}_\delta(\lambda) := \left(\bigcup \{Q \subset [0, 1]^d : f^\delta - \langle f \rangle_Q \leq \lambda\} \right). \quad (2.70)$$

We also need to extend the definition of zooming number [KSU08] 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

$$\tilde{N}_{\delta, \eta, \alpha} := \sup_{\lambda \in \left(\eta^{\frac{1}{d}}, 1 \right]} N_\delta \left((1 + 2\alpha) \log \left(M_d \lambda^d / \eta \right), \lambda \right), \quad (2.71)$$

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 $\mathcal{X}_\delta \left((1 + 2\alpha) \log \left(M_d \lambda^d / \eta \right) \right)$. The number $\tilde{N}_{\delta, \eta, \alpha}$ is well-defined. This is because the $\mathcal{X}_\delta \left((1 + 2\alpha) \log \left(M_d \lambda^d / \eta \right) \right)$ 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.6.

Proposition 3. *Following the Zooming Rule (2.63), we have*

1. *Each parent cube of measure μ is played at most $\frac{2(\Psi + D_\varepsilon)^2 \log(2T^2/\epsilon)}{\alpha^2 [\log(\mu/\eta)]^2}$ episodes.*
2. *Under event $\tilde{\mathcal{E}}_T$, 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 $\tilde{\mathcal{E}}_T$ holds. By Proposition 3, we can regroup the regret. 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 3 (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

$$|\mathcal{K}_i| \leq N_\delta \left((1 + 2\alpha) \log \left(M_d 2^{-di} / \eta \right), 2^{-i} \right), \quad (2.72)$$

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

$$\sum_{t=1}^T \sum_{i=1}^{M_d} \Delta_{t,i}^\delta \leq \sum_{t=1}^T (1 + 2\alpha) M_d \log \left(M_d \mu(Q_t) / \eta \right) \quad (2.73)$$

$$\leq \sum_{i=0}^{I-1} \sum_{Q \in \mathcal{K}_i} \sum_{t \in S_Q} (1 + 2\alpha) M_d \log \left(M_d 2^{-di} / \eta \right), \quad (2.74)$$

where (2.73) uses (2.69), (2.74) regroups the sum as argued above. Using Proposition 3, we can bound (2.74) by:

$$\begin{aligned} & \sum_{i=0}^{I-1} \sum_{Q \in \mathcal{K}_i} \sum_{t \in S_Q} (1 + 2\alpha) M_d \log \left(M_d 2^{-di} / \eta \right) \\ & \leq \sum_{i=0}^{I-1} \sum_{Q \in \mathcal{K}_i} |S_Q| (1 + 2\alpha) M_d \log \left(\frac{M_d 2^{-di}}{\eta} \right) \\ & \stackrel{\textcircled{1}}{\leq} \sum_{i=0}^{I-1} \sum_{Q \in \mathcal{K}_i} \frac{2(\Psi + D_E)^2 \log(2T^2/\epsilon)}{\alpha^2 [\log(2^{-di}/\eta)]^2} \\ & \quad \cdot (1 + 2\alpha) M_d \log \left(\frac{M_d 2^{-di}}{\eta} \right) \end{aligned} \quad (2.75)$$

$$\begin{aligned} & \leq \sum_{i=0}^{I-1} N_\delta \left((1 + 2\alpha) \log \left(M_d 2^{-di} / \eta \right), 2^{-di} \right) \\ & \quad \cdot \frac{2(\Psi + D_E)^2 \log(2T^2/\epsilon)}{\alpha^2 [\log(2^{-di}/\eta)]^2} \cdot (1 + 2\alpha) M_d \log \left(\frac{M_d 2^{-di}}{\eta} \right) \\ & \leq \frac{2(1 + 2\alpha) M_d (\Psi + D_E)^2}{\alpha^2} \tilde{N}_{\delta, \eta, \alpha} \\ & \quad \cdot \log(2T^2/\epsilon) \sum_{i=0}^{I-1} \frac{\log(M_d 2^{-di} / \eta)}{[\log(2^{-di} / \eta)]^2}, \end{aligned} \quad (2.76)$$

where $\textcircled{1}$ uses item 1 in Proposition 3, (2.76) uses (2.72). Recall $\eta = 2^{-dI}$ for some

positive integer I . We can use the above to prove Theorem 6, by using $\eta = 2^{-dI}$ and

$$\begin{aligned}
\sum_{i=0}^{I-1} \frac{\log(M_d 2^{-di}/\eta)}{[\log(2^{-di}/\eta)]^2} &= \sum_{i=0}^{I-1} \frac{\log M_d}{\left[\log \frac{2^{-di}}{\eta}\right]^2} + \sum_{i=0}^{I-1} \frac{1}{\log \frac{2^{-di}}{\eta}} \\
&= \sum_{i=0}^{I-1} \frac{\log M_d}{d^2 (\log 2)^2 (I-i)^2} + \sum_{i=0}^{I-1} \frac{1}{d(\log 2)(I-i)} \\
&= \mathcal{O}(1) + \mathcal{O}(\log I), \\
&= \mathcal{O}(\log \log(1/\eta)),
\end{aligned} \tag{2.77}$$

where the first term in (2.77) is $\mathcal{O}(1)$ since $\sum_{i=1}^{\infty} \frac{1}{i^2} = \mathcal{O}(1)$ and the second term in (2.77) is $\mathcal{O}(\log I)$ by the order of a harmonic sum. The above analysis gives Theorem 6.

Theorem 6. *Choose positive integer I , and let $\eta = 2^{-Id}$. For $\epsilon > 0$ and $t \leq T$, with probability $\geq 1 - 2\epsilon$, for any $\delta > |\mathcal{Q}_t|\eta$ such that δ is f -admissible, Algorithm 4 (with zooming rate α) admits t -episode δ -regret of:*

$$\mathcal{O}\left(\frac{1+2\alpha}{\alpha^2} M_d \Psi^2 \tilde{N}_{\delta,\eta,\alpha} \log\left(\frac{T}{\epsilon}\right) \log \log(1/\eta)\right), \tag{2.78}$$

where $\Psi = \mathcal{O}(\log(T/\epsilon) + \log(1/\eta))$, $\tilde{N}_{\delta,\eta,\alpha}$ is defined in (2.71), 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 6, the definition of $\tilde{N}_{\delta,\eta,\alpha}$ is used in (2.76). For a more refined bound, we can instead use

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

where \hat{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.7.

In Remark 6, we give an example of regret rate on $f(x) = 2 \log \frac{1}{x}$, $x \in (0, 1]$ with specific input parameters.

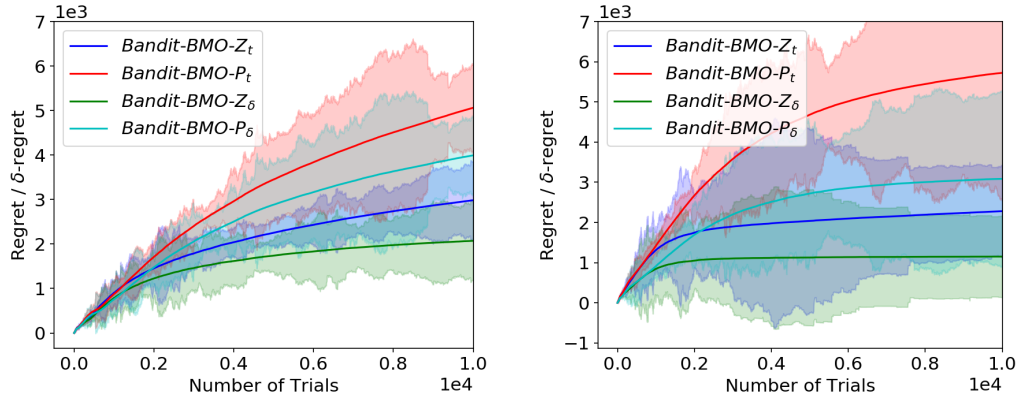


Figure 2.7: Algorithms 3 and 4 on Himmelblau’s function (left) and Styblinski–Tang function (right). Each line is averaged over 10 runs. The shaded area represents one variance above and below the average regret. For the Bandit-BMO-Z algorithm, all arms played incur regret, and each episode has 4 arm trials in it. In the figures, Bandit-BMO-Z $_{\delta}$ (resp. Bandit-BMO-P $_{\delta}$) 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 Figure 2.8.

Remark 6. Consider the (unbounded, BMO) function $f(x) = 2 \log \frac{1}{x}$, $x \in (0, 1]$. Pick $T \geq 20$. For some $t \leq T$, the t -step δ -regret of Algorithm 4 is $\mathcal{O}(\text{poly-log}(t))$ while allowing $\delta = \mathcal{O}(1/T)$ and $\eta = \Theta(1/T^4)$. Intuitively, Algorithm 4 gets close to f^{δ} even if f^{δ} is very large. Details of this example can be found in Appendix A.7.

2.2.6 Experiments

We deploy Algorithms 3 and 4 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 2.7. We measure performance using traditional regret and δ -regret. Traditional regret can be measured because both functions are continuous, in addition to being BMO.

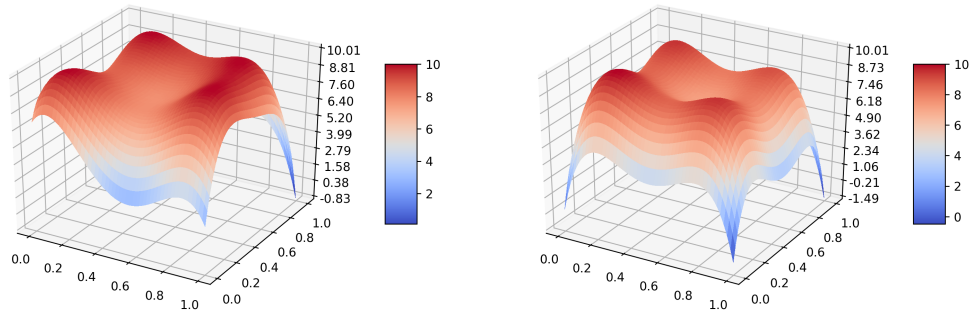


Figure 2.8: Landscapes of test functions used in Section 2.1.3. Left: (Rescaled) Himmelblau’s function. Right: (Rescaled) Styblinski-Tang function.

Discussion on Lower Bounds

A classic trick to derive minimax lower bounds for (stochastic) bandit problems is the “needle-in-a-haystack.” In this argument [Aue02], 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 [KSU08, LWHZ19]. This argument, 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.

2.2.7 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.

Chapter 3

Bandit Learning with Random Walk Feedback

3.1 Introduction

Multi-Armed Bandit (MAB) problems simultaneously call for exploitation of good options and exploration of the decision space. Algorithms for this problem find applications in various domains, from medical trials [Rob52] to online advertisement [LCLS10]. Many authors have studied bandit problems from different perspectives.

In this paper, we study a new bandit learning problem where the feedback is depicted by a random walk over the arms. That is, each time an arm/node i is played, one observes a random walk over the arms/nodes from i to an absorbing node, and the reward is the length of this random walk. Such feedback structure may show up in different scenarios. One concrete motivation is the browsing behavior of internet users within a certain web domain. Specifically, one may view a user's browsing record as a random walk. Web pages within a web domain are viewed as graph nodes. The user transits to another node when she opens another page in this domain. This random walk ends (hits an absorbing node) when the user exits the domain. In this learning setting, we want to carefully select which nodes to initialize the random walks (e.g., recommend an entrance to a web domain), so that the hitting time to the absorbing node (e.g., profit from users' browsing) is maximized. We therefore ask the following question:

In a graph with an absorbing node, if we can select the initial node to seed a random walk and observe the random walk trajectory, how should we select the initial nodes, so that the random walks are as long-lasting as possible? (P)

We study this problem from an online learning perspective. To be more precise, we consider the following model. The environment is modeled by a graph $G = (V, E)$, where V consists of transient nodes $[K] := \{1, 2, \dots, K\}$ and an absorbing node $*$. Each edge ij ($i \in [K], j \in [K] \cup \{*\}$) can encode two quantities, a transition probability from i to j and the distance from i to j . For $t = 1, 2, \dots, T$, we pick a node to start a random walk, and observe the random walk trajectory from the selected node to the absorbing node. For each random walk, we use its hitting time (to the absorbing node $*$) to model how long-lasting it is. With this formulation, we can define a bandit learning problem for the question **(P)**. Each time, the agent picks a node in G to start a random walk, observes the trajectory, and receives the hitting time of the random walk as reward. In this setting, the performance of learning algorithms is typically measured by regret, which is the difference between the rewards of an optimal node and the rewards of the nodes played by the algorithm. Unlike standard multi-armed bandit problems, the feedback is random walk trajectories and thus reveals information not only about the node played, but the environment (transitions/distances among nodes) as well. This new feedback structure calls for new insights on learning with graph random walk feedback.

We start with a stochastic version, where the graph G is fixed and unknown. Intriguingly, we observe that the this problem can be simultaneously be almost as hard as a standard MAB problem, and much easier than a standard MAB problem. More specifically, there exists a difficult problem instance on which the following two facts simultaneously hold: 1. No algorithm can achieve a regret bound independent of problem intrinsic information theoretically; and 2. There exists an algorithm whose performance is independent of problem intrinsic in terms of tail of mistakes. This reveals an intriguing phenomenon in general semi-bandit feedback learning problems.

Then we study an adversarial version, where the edge lengths in graph G change adversarially over time. This setting takes care of changing environments, which can model the potential change of users' preference. We develop a novel variant of the exponential weight algorithm [LW94, ACBFS02] for the adversarial formulation, and provide a new

concentration bound in Lemma 9. In such problems, the adversary does not directly chose reward. Instead, they change rewards by specifying the underlying distribution from which the hitting times are sampled. A high probability regret of order $\tilde{O}\left(\sqrt{\kappa T}\right)$ is proven, where κ depends on the graph structure, instead of number of options.

The lower bound introduces several novel insights compared to regular bandit lower bounds, since our events are from a more difficult sample space. Intuitively, the sample space describes how much information the feedback can carry. If we execute a policy π for T epochs on a problem instance, the sample space is then $\left(\cup_{h=1}^{\infty} \mathcal{B}^h\right)^T$, where \mathcal{B} is the space of all events that a single step on a trajectory can generate. For example, if all edge length can be sampled from $[0, 1]$, then $\mathcal{B} = [0, 1] \cup [K]$, since a single step on a trajectory might be any node, and the corresponding edge length may be any number from $[0, 1]$. In this case the sample space of simple epoch is $\cup_{h=1}^{\infty} \mathcal{B}^h$, where the union up to infinity captures the fact that the trajectory can be arbitrarily long. One can quickly note that this sample space $\left(\cup_{h=1}^{\infty} ([0, 1] \cup [K])^h\right)^T$ is difficult. Indeed, the set $\cup_{h=1}^{\infty} [0, 1]^h$ contains much richer information than of standard MAB problems, which is $[0, 1]^H$ for some finite H . This richer sample space means that: each feedback carries much more information and thus our problem can be strictly easier than a standard MAB. However, the information theoretical lower bounds for our problem are of order $\tilde{\Omega}\left(\sqrt{T}\right)$. In other words, we prove that even though each trajectory carries much more information than a reward sample (and has a chance of revealing all information about the environment when the trajectory is long), no algorithm can beat the bound $\tilde{\Omega}\left(\sqrt{T}\right)$.

In terms of applications, many other scenarios also fit in our model. For example, our model can also describe the browsing over items (e.g., videos, news articles, commodities) in mobile apps. In this case, items are modeled as nodes in a graph, in which tapping a node leads to a transition to another node, and closing the mobile app means hitting an absorbing node. We may also want to prolong the browsing activity in this case. Many other recommendation system applications follow a similar structure.

In summary, our major contributions are

1. We propose a new bandit learning problem motivated by real-world problems with random walk structures, and provide algorithmic solutions to such problems.
2. We observe that in the stochastic setting, this problem can simultaneously be much easier than a standard MAB and as hard as a standard MAB. In fact, a single problem instance can simultaneously display these two properties.
3. We prove lower bounds for this problems in a random walk trajectories sample space. Our results show that, although random walk trajectories carry much more information than reward samples, the additional information does not simplify the problem.
4. We design a new algorithm for the adversarial setting, and provide a high probability bound depending on the graph structure instead of number of options. This bound improves previous counterparts of our adversarial model.

3.1.1 Related Works

Bandit problems date its history back to at least [Tho33], and have been studied extensively in the literature. One of the the most popular approaches to the stochastic bandit problem is the Upper Confidence Bound (UCB) algorithms [Rob52, LR85, Aue02]. Various extensions of UCB algorithms have been studied [SKKS10a, AYPS11, AG12, BS12, SS14]. Specifically, some works use KL-divergence to construct the confidence bound [LR85, GC11, MMS11], or include variance estimates within the confidence bound [AMS09, AO10]. UCB is also used in the contextual learning setting (e.g., [LCLS10, KO11, Sli14]). Parallel to the stochastic setting, studies on the adversarial bandit problem form another line of literature. Since randomized weighted majorities [LW94], exponential weights remains a top strategy for adversarial bandits [ACBFS95, CBFH⁺97, ACBFS02]. Many efforts have been made to improve/extend exponential algorithms. For example, [KNVM14] target at implicit variance reduction. [MS11, ACBGM13] study a partially observable setting. Despite the large body of literature, no previous work has, to the best of our knowledge, explicitly focused the question (**P**). Specifically, if one applies vanilla bandit algorithms without using the

graph structure. The regret bound would depend on the number of options (arms/nodes). This may be much worse than an environment-dependent bound.

For both stochastic bandits and adversarial bandits, lower bounds in different scenarios have been derived, since the $\mathcal{O}(\log T)$ asymptotic lower bounds for consistent policies [LR85]. Worst case bound of order $\mathcal{O}(\sqrt{T})$ have also been derived [ACBFS95] for the stochastic setting. In addition to the classic stochastic setting, lower bounds in other stochastic (or stochastic-like) settings have also been considered, including PAC-learning complexity [MT04], best arm identification complexity [KCG16, CLQ17], and lower bounds in continuous spaces [KSU08]. Lower bound problems for adversarial bandits may be converted to lower bound problems for stochastic bandits [ACBFS95] in many cases. An intriguing lower bound beyond the expected regret is the high probability lower bound of order $\tilde{\mathcal{O}}(\sqrt{T})$ by [GL16].

For the adversarial setting, the Stochastic Shortest Path (with adversarial edge length) [BT91, RM20] and the online MDP problems [EDKM09, GNSA10, DGS14, JLL⁺19] are related to our problem. However, our algorithm achieves better regret rate than the previous results. In particular, our regret bound depends on the connectivity of the transition graph, instead of number of nodes/states/arms.

Another related setting is bandit with side information [MS11, ACBDK15, ACBG⁺17]. In such problems, playing a node reveals information about other nodes, where the feedback structure is governed by an observation graph. Such problem assumes that the observation graph is revealed, either before or after the player has made a decision. In our setting, however, the observation model is unknown and needs to be learned. Very importantly, our setting has much more randomness than that for bandit with side information and is harder in a statistical sense; See Section 3.4 for more discussion.

While bandit problems have been studied in different settings using various techniques, no prior works, to the best of our knowledge, focus on answering the important question **(P)**. Our paper provides a comprehensive answer to this important class of problems **(P)** for propagation over graphs.

3.2 Problem Setting

In this section, we formulate the problem and put forward notations and definitions that will be used throughout the rest of the paper. The learning process repeats for T epochs and the learning environment is described by graphs G_1, G_2, \dots, G_T for epochs $t = 1, 2, \dots, T$. The graph G_t is defined on K transient nodes $[K] = \{1, 2, \dots, K\}$ and one absorbing node denoted by $*$. We will use $V = [K]$ to denote the set of transient nodes, and use $\tilde{V} := [K] \cup \{*\}$ to denote the transient nodes together with the absorbing node. On this node set \tilde{V} , graph G_t encodes transition probabilities and edge lengths: $G_t := \left(\{m_{ij}\}_{i \in V, j \in \tilde{V}}, \{l_{ij}^{(t)}\}_{i \in V, j \in \tilde{V}} \right)$, where m_{ij} is the probability of transiting from i to j and $l_{ij}^{(t)} \in [0, 1]$ is the length from i to j (at epoch t). We gather the transition probabilities among transient nodes to form a transition matrix $M = [m_{ij}]_{i, j \in [K]}$. We make the following assumption about M .

Assumption 2. *The transition matrix $M = [m_{ij}]_{i, j \in [K]}$ among transient nodes is primitive.¹ In addition, there is a constant ρ , such that $\|M\|_\infty \leq \rho < 1$, where $\|M\|_\infty = \max_{i \in [K]} \sum_{j \in [K]} |m_{ij}|$ is the maximum absolute row sum.*

In Assumption 2, the primitivity assumption ensures that we can get to any transient node v from any other node state u . The infinite norm of M being strictly less than 1 means that the random walk will transit to the absorbing node starting from any node (eventually with probability 1). This describes the absorptiveness of the environment. Note that this infinite norm assumption can be replaced by other notions of matrix norms.

Playing node j at epoch t generates a random walk trajectory $\mathcal{P}_{t,j} := (X_{t,0}^{(j)}, L_{t,1}^{(j)}, X_{t,1}^{(j)}, L_{t,2}^{(j)}, X_{t,2}^{(j)}, \dots, L_{t,H_{t,j}}^{(j)}, X_{t,H_{t,j}}^{(j)})$, where $X_{t,0}^{(j)} = j$ is the starting nodes, $X_{t,H_{t,j}}^{(j)} = *$ is the absorbing node, $X_{t,i}^{(j)}$ is the i -th node in the random walk trajectory, $L_{t,i}^{(j)}$ is the edge length from $X_{t,i-1}^{(j)}$ to $X_{t,i}^{(j)}$, and $H_{t,j}$ is the number of edges in trajectory $\mathcal{P}_{t,j}$. For simplicity, we write $X_{t,i}^{(j)}$ (resp. $L_{t,i}^{(j)}$) as $X_{t,i}$ (resp. $L_{t,i}$) when it is clear from context.

For the random trajectory $\mathcal{P}_{t,j} := (X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \dots, L_{t,H_{t,j}}, X_{t,H_{t,j}})$, the

¹A matrix M is primitive if there exists a positive integer k , such that all entries in M^k is positive.

length of the trajectory (or **hitting time** of node j at epoch t) is defined as

$$\mathcal{L}(\mathcal{P}_{t,j}) := \sum_{i=1}^{H_{t,j}} L_{t,i}. \quad (3.1)$$

Here we use the edge length to represent the reward of the trajectory. In practice, the edge lengths may have real-world meanings. For example, the out-going edge from a node may represent utility (e.g., profit) of visiting this node. At epoch t , the agent selects a node $J_t \in [K]$ to initiate a random walk, and observe trajectory \mathcal{P}_{t,J_t} . In stochastic environments, the environment does not change across epochs. Thus for any fixed node $v \in [K]$, the random trajectories $\mathcal{P}_{1,v}, \mathcal{P}_{2,v}, \mathcal{P}_{3,v}, \dots$ are independently identically distributed.

We also define a notion of centrality that will be used later.

Definition 5. Let $X_0, X_1, X_2, \dots, X_\tau = *$ be nodes on a random trajectory. Under Assumption 2, we define, for node $v \in [K]$,

$$\alpha_v := \min_{u \in [K], u \neq v} \mathbb{P}(v \in \{X_1, X_2, \dots, X_\tau\} | X_0 = u)$$

to be the **hitting centrality** of node v . We also define $\alpha = \min_v \alpha_v$

Hitting centrality of a node v is how likely it is visited by a trajectory starting from another node. In non-absorptive (and ergodic) Markov chains, the hitting centrality of any node is 1. This quantity is less than 1 for networks with absorbing nodes. The hitting centralities describe the connectivity of the transition graph.

3.3 Stochastic Setting

In the stochastic setting, the graphs G_t do not change across epochs. To solve this problem, one can estimate the expected hitting times $\mu_j := \mathbb{E}[\mathcal{L}(\mathcal{P}_{t,j})]$ for all nodes $j \in [K]$ (and maintain a confidence interval of the estimations). As one can expect, the random walk trajectory reveals more information than a sample of reward. Naturally, this allows us to reduce this problem to a standard (stochastic) MAB problem.

3.3.1 Reduction to Standard MAB

Recall $\mathcal{P}_{t,J_t} = (X_{t,0}, L_{t,1}, X_{t,1}, \dots, L_{t,H_t,J_t}, X_{t,H_t,J_t})$ is the trajectory at epoch t . For a node v and the trajectories $\mathcal{P}_{1,J_1}, \mathcal{P}_{2,J_2}, \mathcal{P}_{3,J_3}, \dots$, let $k_{v,i}$ be the index (epoch) of the i -th trajectory that covers node v . Let $Y_{v,k_{v,i}}$ be the sum of edge lengths between the first occurrence of v and the absorbing node $*$ in trajectory $k_{v,i}$. One has the following proposition due to Markov property.

Proposition 4. *In the stochastic setting, for any nodes $v \in [K]$, we have, for $\forall t, i \in \mathbb{N}_+, \forall r \in \mathbb{R}$*

$$\mathbb{P}(Y_{v,k_{v,i}} = r) = \mathbb{P}(\mathcal{L}(\mathcal{P}_{t,v}) = r). \quad (3.2)$$

Proof. In a trajectory $\mathcal{P}_{t,J_t} = (X_{t,0}, L_{t,1}, X_{t,1}, \dots, L_{t,H_t,J_t}, X_{t,H_t,J_t})$, conditioning on $X_{t,i} = j$ being known (and no future information is revealed), the randomness generated by $L_{t,i+1}, X_{t,i+1}, L_{t,i+2}, X_{t,i+2}, \dots$ is identical to the randomness generated by $L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \dots$ conditioning on $X_{t,0} = j$ being fixed. Note that even if each trajectory can visit a node multiple times, only one hitting time sample can be used. This is because extracting multiple sample would break Markovianity, by revealing that the random walk will visit a same node again. \square

For a node $v \in [K]$, we define

$$N_t(v) := 1 \vee \sum_{s < t} \mathbb{I}_{[J_s=v]}, \quad N_t^+(v) := 1 \vee \sum_{s < t} \mathbb{I}_{[v \in \mathcal{P}_{s,J_s}]}. \quad (3.3)$$

where $a \vee b = \max\{a, b\}$. In (3.3), $N_t(v)$ is the number of times node v is played, and $N_t^+(v)$ is the number of times node v is covered by a trajectory.

As Proposition 4 suggests, number of times a node is visited linearly accumulates with number of epochs t . We state this observation below in Lemma 5.

Lemma 5. *For any $v \in V$ and a positive integer t*

$$\mathbb{P}\left(N_t^+(v) - N_t(v) - \alpha_v(t - N_t(v)) \geq -\lambda\right) \leq \exp\left(-\frac{\lambda^2}{2t}\right). \quad (3.4)$$

Proof. Recall $\mathcal{P}_{t,J_t} = (X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \dots, L_{t,H_t,J_t}, X_{t,H_t,J_t})$ is the trajectory for epoch t and $X_{i,0}$ is the node played at epoch i . For a fixed node $v \in V$, consider the random variables $\left\{ \mathbb{I}_{[v \in \mathcal{P}_{t,J_t} \setminus \{X_{t,0}\}]} \right\}_t$, which is the indicator that takes value 1 when v is covered in path \mathcal{P}_{t,J_t} but is not played at t . From this definition, we have

$$\sum_{k=1}^t \mathbb{I}_{[v \in \mathcal{P}_{k,J_k} \setminus \{X_{k,0}\}]} = N_t^+(v) - N_t(v).$$

From definition of α_v , we have

$$\mathbb{E} [N_t^+(v) - N_t(v)] = \mathbb{E} \left[\sum_{k=1}^t \mathbb{I}_{[v \in \mathcal{P}_k \setminus \{X_{k,0}\}]} \right] \geq \alpha_v (t - \mathbb{E} [N_t(v)]).$$

Thus by one-sided Azuma's inequality, we have for any $\lambda > 0$,

$$\mathbb{P} \left(N_t^+(v) - N_t(v) - \alpha_v (t - N_t(v)) \geq -\lambda \right) \leq \exp \left(-\frac{\lambda^2}{2t} \right). \quad (3.5)$$

□

By Proposition 4 and Lemma 5, the information about the node rewards (hitting time to absorbing node) accumulates linearly as we play. Thus solving the problem is not hard: one can extract the hitting time estimates and deploy a UCB algorithm based on it. However, some information is lost when we exact hitting time samples, since trajectories also carry additional information (e.g, about graph transition) but we only exactly hitting time samples. Thus the intriguing questions to ask are:

1. How much easier is this problem than its standard MAB counterpart? **(Q1)**
2. Is a reduction to standard MAB optimal? Do we give up too much information by only extracting hitting time samples from trajectories? **(Q2)**

The answers to **(Q1)** and **(Q2)** are intriguing as we show in Sections 3.3.2 and 3.3.2. In fact, the problem can be much easier than a standard MAB and as hard as a standard MAB at the same time (Theorems 7 and 8), and exacting hitting time samples does not give up critical information (Theorem 8).

3.3.2 Is this Problem Much Easier than Standard MAB?

We compare our problem with the standard MAB, in a setting where the time horizon T is fixed and the transition graph is loosely connected. In such situations, two seemingly contradictory facts can simultaneously hold: 1. the problem is significantly easier than the standard MAB, particularly in a distributional sense. 2. the problem can be as hard as the standard MAB problem information-theoretically. Below, we study item 1 in Section 3.3.2, study item 2 in Section 3.3.2, and highlight the intriguing observations in Remark 7. Also, an answer to **(Q2)** is provided via Theorem 8 in Section 3.3.2, which is, no critical information is lost even if we only exact hitting time samples.

This Problem Can Be Much Easier than Standard MAB

Our problem is easier than the standard MAB problem, not only in terms of expected regret, but also in a distributional sense. More specifically, the “tail-of-mistakes” (Theorem 7) of the UCB algorithm decays very fast, even if the connection is loose. Next, we state the UCB algorithm using our notations, and present a new tail-of-mistakes bound.

For a transient node $v \in [K]$ and n trajectories (at epochs $k_{v,1}, k_{v,2}, \dots, k_{v,n}$) that cover node v , the hitting time estimator of v is computed as

$$\tilde{Z}_{v,n} := \frac{1}{n} \sum_{i=1}^n Y_{v,k_{v,i}}. \quad (3.6)$$

Since $v \in \mathcal{P}_{t,v}$, $Y_{v,k_{v,i}}$ is an identical copy of the hitting time $\mathcal{L}(\mathcal{P}_{t,v})$ (Proposition 4). For (3.6), one can also use robust mean estimators [BCBL13], but a plain mean estimator is sufficient for the purpose of this paper.

We also need confidence intervals for our estimators. Given $N_t^+(v)$ trajectories covering v , the confidence terms (at epoch t) are

$$\tilde{C}_{N_t^+(v),t} := \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}},$$

where $\xi_t = \max \left\{ 1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log \rho} + \frac{5 \log t}{\log 1/\rho} \right\}$.

At each time t , we play a node J_t that maximizes the UCB index, $\tilde{Z}_{v, N_t^+(v)} + \tilde{C}_{N_t^+(v), t}$. This strategy is described in Algorithm 5. In a distributional sense, the UCB algorithm's

Algorithm 5

- 1: **Input:** A set of nodes $[K]$. Parameters: a constant ρ that bounds the spectral radius of M .
- 2: **Warm up:** Play each node once to initialize. Observe trajectories.
- 3: For any $v \in [K]$, define the decision index

$$I_{v, N_t^+(v), t} = \tilde{Z}_{v, N_t^+(v)} + \tilde{C}_{N_t^+(v), t}. \quad (3.7)$$

- 4: **for** $t = 1, 2, 3, \dots$ **do**

- 5: Select J_t to start a random walk, such that

$$J_t \in \arg \max_{v \in V} I_{v, N_t^+(v), t}, \quad (3.8)$$

with ties broken arbitrarily.

- 6: Observe the trajectory $\mathcal{P}_{t, v_t} := \{X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \dots, L_{t, H_{t, v_t}}, X_{t, H_{t, v_t}}\}$.
 Update $N_t^+(\cdot)$ and decision indices for all $v \in [K]$.
-

tail decay very fast, even if α_v is only slightly positive. This is summarized in Theorem 7.

Theorem 7 (Tail of Mistakes). *Fix any $\beta \in (0, \frac{1}{2})$. For a sub-optimal node v , define $K_{\min, v} := \min_{t \in \mathbb{N}} \left\{ \sqrt{\frac{8\xi_t \log t}{\alpha_v t - t^{\frac{1}{2} + \beta}}} \leq \frac{1}{2} \Delta_v \right\}$, where $\Delta_v := \min_{j \in [K]} \mu_j - \mu_v$ is the optimality gap. For any sub-optimal node v , and $x, T \geq K_{\min, v}$, we have*

$$\mathbb{P}(N_T(v) \geq x) \lesssim \frac{1}{(1 + \alpha_v)x^3} + \frac{\sqrt{x}}{\Delta_v \sqrt{\alpha_v \log 1/\rho}} \exp\left(-\Delta_v \sqrt{\alpha_v x \log 1/\rho}\right) + Kx \exp\left(-x^{2\beta}\right). \quad (3.9)$$

In particular, one has $\mathbb{P}(N_T(v) \geq x) \lesssim \frac{1}{(1 + \alpha_v)x^3} \leq \frac{1}{2x^3}$ by omitting terms of exponentially smaller order.

Proof. For any positive integer u , and $\beta \in (0, \frac{1}{2})$, consider event

$$\mathcal{E}_{u,\beta} := \left\{ N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -t^{\frac{1}{2}+\beta}, \text{ for all } t \geq u \text{ and all } v \in V \right\}.$$

By Lemma 5 and a union bound, we have $\mathbb{P}(\mathcal{E}_{u,\beta}) \geq 1 - K \sum_{t=u}^{\infty} \exp(-u^{2\beta}) \gtrsim 1 - Kx \exp(-x^{2\beta})$. Under event $\mathcal{E}_{u,\beta}$, at any time $t \in [u, T]$, we have $N_t^+(v) \geq \alpha_v t - t^{\frac{1}{2}+\beta}$ for all $v \in V$.

We define an ‘‘alternative’’ index at time t by:

$$I'_{v,t} := \tilde{Z}_{v,N_t^+(v)} + \sqrt{\frac{8\xi_t \log t}{\alpha_v t - t^{\frac{1}{2}+\beta}}}. \quad (3.10)$$

By Lemma 5, we know $I'_{v,t} \geq I_{v,N_t^+(v),t}$ with high probability.

Fix $\beta \in (0, \frac{1}{2})$. For any $\tau \in \mathbb{R}$, any positive integer w , and any sub-optimal node v , we have

$$\left\{ N_T(v) \leq w \right\} \Leftarrow \left\{ I'_{v,t} \leq \tau \text{ for all } t \in [w, T] \right\} \cap \mathcal{E}_{w,\beta} \cap \left\{ I_{v^*,s,s+w} > \tau \text{ for all } s \in [\alpha_v w - w^{\frac{1}{2}+\beta}, T] \right\}. \quad (3.11)$$

This is because (3.11) ensures that after time w , the index of node v is always lower than the index of node v^* .

By taking the contrapositive of (3.11), we have, for arbitrary $\tau \in \mathbb{R}$,

$$\{N_T(v) > w\} \Rightarrow \left\{ \exists t \in [w, T] \text{ s.t. } I'_{v,t} > \tau \right\} \cup \left\{ \exists s \in [\alpha_v w - w^{\frac{1}{2}+\beta}, T] \text{ s.t. } I_{v^*,s,s+w} \leq \tau \right\} \cup \overline{\mathcal{E}_{w,\beta}}. \quad (3.12)$$

By taking probability on both sides of (3.12), we get, for any $\tau \in \mathbb{R}$ and $w \in \mathbb{N}$, we have

$$\mathbb{P}(N_T(v) > w) \leq \sum_{t=w}^T \mathbb{P}(I'_{v,t} > \tau) + \sum_{s=\alpha_v w - w^{\frac{1}{2}+\beta}}^T \mathbb{P}(I_{v^*,s,w+s} \leq \tau) + \mathbb{P}(\overline{\mathcal{E}_{w,\beta}}).$$

By taking $\tau = \mathbb{E}[Z_{v^*}]$, and $w = x$, we get

$$\mathbb{P}(N_T(v) > x) \lesssim \sum_{t=x}^T \mathbb{P}(I'_{v,t} > \mathbb{E}[Z_{v^*}]) + \sum_{s=\alpha_v x - x^{\frac{1}{2}+\beta}}^T \mathbb{P}(I_{v^*,s,w+s} \leq \mathbb{E}[Z_{v^*}]) + Kx \exp(-x^{2\beta}). \quad (3.13)$$

For $t \geq x \geq K_{\min, v}$, we have

$$\sqrt{\frac{8\xi_t \log t}{\alpha_v t - t^{\frac{1}{2} + \beta}}} \leq \frac{1}{2} \Delta_v. \quad (3.14)$$

Under event $\mathcal{E}_{w, t}$, we have $N_t^+(v) \geq \alpha_v t - t^{\frac{1}{2} + \beta}$. Thus we have

$$\begin{aligned} \mathbb{P}(I'_{v, t} > \mathbb{E}[Z_{v^*}]) &= \mathbb{P}\left(\tilde{Z}_{v, N_t^+(v)} + \sqrt{\frac{8\xi_t \log t}{\alpha_v t - t^{\frac{1}{2} + \beta}}} > \mathbb{E}[Z_v] + \Delta_v\right) \\ &\leq \mathbb{P}\left(\tilde{Z}_{v, N_t^+(v)} > \mathbb{E}[Z_v] + \frac{1}{2}\Delta_v\right) \\ &\stackrel{\textcircled{1}}{\leq} \max\left\{3 \exp\left(-\frac{\Delta_v^2(1-\rho)^3}{8\rho} N_t^+(v)\right), 3 \exp\left(-\Delta_v \sqrt{\frac{N_t^+(v) \log 1/\rho}{10}}\right)\right\} \\ &\stackrel{\textcircled{2}}{\lesssim} \max\left\{3 \exp\left(-\frac{\Delta_v^2(1-\rho)^3}{8\rho} \alpha_v t\right), 3 \exp\left(-\Delta_v \sqrt{\frac{\alpha_v t \log 1/\rho}{10}}\right)\right\}, \end{aligned}$$

where $\textcircled{1}$ uses Lemma 14 (in Appendix B.1.1), and $\textcircled{2}$ again uses Lemma 5.

The above gives

$$\begin{aligned} \sum_{t=x}^{\infty} \mathbb{P}(I'_{v, t} > \mathbb{E}[Z_{v^*}]) &\lesssim \sum_{t=x}^{\infty} \max\left\{3 \exp\left(-\frac{\Delta_v^2(1-\rho)^3}{8\rho} \alpha_v t\right), 3 \exp\left(-\Delta_v \sqrt{\frac{\alpha_v t \log 1/\rho}{10}}\right)\right\} \\ &\lesssim \frac{\sqrt{x}}{\Delta_v \sqrt{\alpha_v \log 1/\rho}} \exp\left(-\Delta_v \sqrt{\alpha_v x \log 1/\rho}\right). \end{aligned} \quad (3.15)$$

Also by Lemma 14, we have

$$\mathbb{P}(I_{v^*, s, s+u} \geq \mathbb{E}[Z_{v^*}]) = \mathbb{P}\left(\tilde{Z}_{v^*, s} + \tilde{C}_{s, s+u} \geq \mathbb{E}[Z_{v^*}]\right) \leq (u+s)^{-4}.$$

This gives

$$\begin{aligned} \sum_{s=\alpha_v x - x^{2\beta}}^{\infty} \mathbb{P}(I_{v^*, s, s+u} \geq \mathbb{E}[Z_{v^*}]) &= \sum_{s=\alpha_v x - x^{2\beta}}^{\infty} \mathbb{P}\left(\tilde{Z}_{v^*, s} + \tilde{C}_{s, s+u} \geq \mathbb{E}[Z_{v^*}]\right) \\ &\leq \sum_{s=\alpha_v x - x^{2\beta}}^{\infty} (x+s)^{-4} \\ &\lesssim \frac{1}{(1+\alpha_v)x^3}. \end{aligned} \quad (3.16)$$

Collecting terms from (3.13), (3.15), (3.16), and rearranging concludes the proof. \square

This Problem Can Be as Hard as Standard MAB

The problem can be at the same time as hard as a standard MAB, as noted below in Theorem 8. Also, this result gives an answer to **(Q2)**: Even if only extracting hitting time samples loses information, no algorithm can perform better than simply extracting hitting time samples.

Theorem 8. *For any given T and any policy π , there exists a problem instance \mathfrak{J} satisfying Assumption 2 such that, for any $\epsilon, \delta \in (0, \frac{1}{4})$, the T step regret of π on instance \mathfrak{J} is lower bounded by*

$$(2\epsilon + O(\epsilon^2) + O(\epsilon\delta^2)) T \exp(-8T(\epsilon^2 + O(\epsilon^3) + O(\epsilon^2\delta))). \quad (3.17)$$

In particular, setting $\epsilon = \frac{1}{4}T^{-1/2}$ and $\delta = O(\epsilon^{1/2})$ gives that there exists a problem instance such that the regret of any algorithm on this instance is lower bounded by $\Omega(\sqrt{T})$.

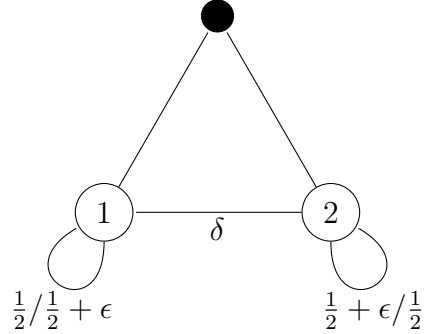


Figure 3.1: Problem instances constructed to prove Theorem 8. The edge labels denote edge transition probabilities in $\mathfrak{J}/\mathfrak{J}'$.

Proof. We construct two “symmetric” problem instances \mathfrak{J} and \mathfrak{J}' both on two transient nodes $\{1, 2\}$ and one absorbing node $*$. All edges in both instances are of length 1. We use $M = [m_{ij}]$ (resp. $M' = [m'_{ij}]$) to denote the transition probabilities among transient nodes in instance \mathfrak{J} (resp. \mathfrak{J}'). We construct instances \mathfrak{J} and \mathfrak{J}' so that $M = \begin{bmatrix} \frac{1}{2} & \delta \\ \delta & \frac{1}{2} + \epsilon \end{bmatrix}$

and $M' = \begin{bmatrix} \frac{1}{2} + \epsilon & \delta \\ \delta & \frac{1}{2} \end{bmatrix}$, as shown in Figure 3.1. In other words, M' is the anti-transpose (transpose with respect to the anti-diagonal) of M .

We use Z_v (resp. Z'_v) to denote the random variable $\mathcal{L}(\mathcal{P}_{t,v})$ in problem instance \mathfrak{J} (resp. \mathfrak{J}'). Then we have

$$\begin{bmatrix} \mathbb{E}[Z_1] \\ \mathbb{E}[Z_2] \end{bmatrix} = M \begin{bmatrix} \mathbb{E}[Z_1] \\ \mathbb{E}[Z_2] \end{bmatrix} + \mathbf{1}, \quad \begin{bmatrix} \mathbb{E}[Z'_1] \\ \mathbb{E}[Z'_2] \end{bmatrix} = M' \begin{bmatrix} \mathbb{E}[Z'_1] \\ \mathbb{E}[Z'_2] \end{bmatrix} + \mathbf{1},$$

where $\mathbf{1}$ is the all-one vector. Solving the above equations gives, for both instances \mathfrak{J} and \mathfrak{J}' , the optimality gap Δ is

$$\Delta := |\mathbb{E}[Z_1] - \mathbb{E}[Z_2]| = 2\epsilon + O(\epsilon^2) + O(\epsilon\delta^2). \quad (3.18)$$

Let π be any fixed algorithm and let T be any fixed time horizon, we use $\mathbb{P}_{\mathfrak{J},\pi}$ (resp. $\mathbb{P}_{\mathfrak{J}',\pi}$) to denote the probability measure of running π on instance \mathfrak{J} (resp. \mathfrak{J}') for T epochs.

Since the event $\{J_t = 1\}$ ($t \leq T$) is measurable by both $\mathbb{P}_{\mathfrak{J},\pi}$ and $\mathbb{P}_{\mathfrak{J}',\pi}$, by the Bretagnolle-Huber inequality we have

$$\mathbb{P}_{\mathfrak{J},\pi}(J_t = 0) + \mathbb{P}_{\mathfrak{J}',\pi}(J_t = 1) \geq \frac{1}{2} \exp(-D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \parallel \mathbb{P}_{\mathfrak{J}',\pi})) \quad (3.19)$$

where we use Pinsker's inequality for the last inequality.

Let \mathcal{Q}_i (resp. \mathcal{Q}'_i) be the probability measure generated by playing node i in instance \mathfrak{J} (resp. \mathfrak{J}'). We can then decompose $\mathbb{P}_{\mathfrak{J},\pi}$ by

$$\begin{aligned} \mathbb{P}_{\mathfrak{J},\pi} &= \mathcal{Q}_{J_1} \mathbb{P}(J_1|\pi) \mathcal{Q}_{J_2} \mathbb{P}(J_2|\pi, J_1) \cdots \mathcal{Q}_{J_T} \mathbb{P}(J_T|\pi, J_1, J_2, \dots, J_{t-1}), \\ \mathbb{P}_{\mathfrak{J}',\pi} &= \mathcal{Q}'_{J_1} \mathbb{P}(J_1|\pi) \mathcal{Q}'_{J_2} \mathbb{P}(J_2|\pi, J_1) \cdots \mathcal{Q}'_{J_T} \mathbb{P}(J_T|\pi, J_1, J_2, \dots, J_{t-1}). \end{aligned}$$

By chain rule for KL-divergence, we have

$$\begin{aligned} D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \parallel \mathbb{P}_{\mathfrak{J}',\pi}) &= \sum_{J_1 \in \{1,2\}} \mathbb{P}(J_1|\pi) D_{KL}(\mathcal{Q}_{J_1} \parallel \mathcal{Q}'_{J_1}) \\ &\quad + \sum_{t=2}^T \sum_{J_t \in \{1,2\}} \mathbb{P}(J_t|\pi, J_1, \dots, J_{t-1}) D_{KL}(\mathcal{Q}_{J_t} \parallel \mathcal{Q}'_{J_t}). \end{aligned} \quad (3.20)$$

Since the policy must pick one of node 1 and node 2, from nonnegativity of KL-divergence and (3.20) we have

$$D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \|\mathbb{P}_{\mathfrak{J}',\pi}) \leq \sum_{t=1}^T \sum_{i=1}^2 D_{KL}(\mathcal{Q}_i \|\mathcal{Q}'_i), \quad (3.21)$$

which allows us to remove dependence on policy π .

Next we study the distributions \mathcal{Q}_i . With edge lengths fixed, the sample space of this distribution is $\cup_{h=1}^{\infty} \{1, 2\}^h$, since length of the trajectory can be arbitrarily long, and each node on the trajectory can be either of $\{1, 2\}$. To describe the distribution \mathcal{Q}_i and \mathcal{Q}'_i , we use random variables $X_0, X_1, X_2, X_3, \dots \in \{1, 2, *\}$ (with $X_0 = i$), where X_k is the k -th node in the trajectory generated by playing i .

By Markov property we have, for $i, j \in \{1, 2\}$,

$$\mathcal{Q}_i(X_{k+1}, X_{k+2}, \dots | X_k = j) = \mathcal{Q}_j \quad \text{and} \quad \mathcal{Q}'_i(X_{k+1}, X_{k+2}, \dots | X_k = j) = \mathcal{Q}'_j, \quad \forall k \in \mathbb{N}_+.$$

In words, conditioning on k -th node being j , the distribution generated by subsequent nodes are the same as \mathcal{Q}_j .

Note we can decompose \mathcal{Q}_i by $\mathcal{Q}_i = \mathcal{Q}_i(X_1)\mathcal{Q}_i(X_2, X_3, \dots, |X_1)$. Thus by chain rule of KL-divergence, for $i \in \{1, 2\}$, and $j \neq i$,

$$\begin{aligned} & D_{KL}(\mathcal{Q}_i \|\mathcal{Q}'_i) \\ &= D_{KL}(\mathcal{Q}_i(X_1) \|\mathcal{Q}'_i(X_1)) \\ & \quad + \sum_{x_1 \in \{1, 2\}} \mathcal{Q}_i(X_1 = x_1) D_{KL}(\mathcal{Q}_i(X_2, X_3, \dots | X_1 = x_1) \|\mathcal{Q}'_i(X_2, X_3, \dots | X_1 = x_1)) \\ &= D_{KL}(\mathcal{Q}_i(X_1) \|\mathcal{Q}'_i(X_1)) + \mathcal{Q}_i(X_1 = i) D_{KL}(\mathcal{Q}_i \|\mathcal{Q}'_i) + \mathcal{Q}_i(X_1 = j) D_{KL}(\mathcal{Q}_j \|\mathcal{Q}'_j), \quad (3.22) \end{aligned}$$

where the last step uses the Markov property.

Since $D_{KL}(\mathcal{Q}_i(X_1) \|\mathcal{Q}'_i(X_1)) = 2\epsilon^2 + O(\epsilon^3) + O(\epsilon^2\delta)$ for $i \in \{1, 2\}$, the above (Eq. 3.22) gives

$$D_{KL}(\mathcal{Q}_i \|\mathcal{Q}'_i) = 4\epsilon^2 + O(\epsilon^3) + O(\epsilon^2\delta). \quad (3.23)$$

Combining (3.21) and (3.23) gives

$$D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \|\mathbb{P}_{\mathfrak{J}',\pi}) \leq 8T(\epsilon^2 + O(\epsilon^3) + O(\epsilon^2\delta)). \quad (3.24)$$

Let $\text{Reg}(T)$ (resp. $\text{Reg}'(T)$) be the T epoch regret in instance \mathfrak{J} (resp. \mathfrak{J}').

Recall, by our construction, node 1 is suboptimal in instance \mathfrak{J} and node 1 is optimal in instance \mathfrak{J}' . Since the optimality gaps in \mathfrak{J} and \mathfrak{J}' are the same (Eq. 3.18), we have,

$$\begin{aligned} \text{Reg}(T) + \text{Reg}'(T) &\geq \Delta \sum_{t=1}^T \left(\mathbb{P}_{\mathfrak{J},\pi}(J_t = 1) + \mathbb{P}_{\mathfrak{J}',\pi}(J_t = 0) \right) \\ &\geq \frac{1}{2} \Delta T \exp(-D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \|\mathbb{P}_{\mathfrak{J}',\pi})) && \text{(by Eq. 3.19)} \\ &\geq \frac{1}{2} \Delta T \exp(-8T(\epsilon^2 + O(\epsilon^3) + O(\epsilon^2\delta))) && \text{(by Eq. 3.24)} \\ &\geq (2\epsilon + O(\epsilon^2) + O(\epsilon\delta^2)) T \exp(-8T(\epsilon^2 + O(\epsilon^3) + O(\epsilon^2\delta))). && \text{(by Eq. 3.18)} \end{aligned}$$

Now, we set $\epsilon = \frac{1}{4}T^{-1/2}$ and $\delta = O(\epsilon^{1/2})$, and get $\exp(-8T(\epsilon^2 + O(\epsilon^3) + O(\epsilon^2\delta))) \gtrsim O(1)$. In this case,

$$\text{Reg}(T) + \text{Reg}'(T) \gtrsim \frac{1}{2}\sqrt{T} + O(1),$$

which concludes the proof. □

We explain in Remark 7 why the above results (Theorems 7 and 8) are intriguing.

Remark 7. *Pick a small number $\epsilon > 0$ and fix a time horizon $T = \Theta(\frac{1}{\epsilon^4})$. Consider the instance shown in Figure 3.1, with $\delta = \epsilon^{1/2}$. By Theorem 8, in the first \sqrt{T} epochs, no algorithm can achieve a regret rate better than $\Omega(T^{1/4}) = \Omega(\frac{1}{\epsilon})$. Thus the overall T epoch regret is at least $\Omega(\frac{1}{\epsilon})$. At the same time, the conditions in Theorem 7 is satisfied. More specifically, $\sqrt{\frac{8\xi_T \log T}{\alpha_v T - T^{\frac{1}{2} + \beta}}} \leq \frac{1}{2}\Delta_v$ for some $\beta \in (0, \frac{1}{2})$, since $\alpha_v = \Theta(\delta)$, $\Delta_v = \Theta(\epsilon)$. Thus we can apply Theorem 7 and get, for the suboptimal node v , $\mathbb{P}(N_T(v) \geq x) \lesssim \frac{1}{(1+\alpha_v)x^3} \leq \frac{1}{2x^3}$ for all $x \geq K_{\min,v}$ (defined in Theorem 7). Note that this tail of mistakes does not depend on ϵ (or T), whereas the lower bound does.*

This observation highlights a previously unnoticed phenomenon for semi-bandit feedback problems. There exists a difficult problem instance such that, simultaneously on this instance, 1. no algorithm can achieve a regret rate independent of the problem instance or the time horizon in the information theoretical sense, and 2. the UCB algorithm's tail of mistake is independent of problem instance and the time horizon.

3.3.3 Regret Analysis

As discussed previously, the problem with random walk feedback can be reduced to standard MAB problems, and the UCB algorithm solves this problem as one would expect. We present here the regret upper bound for algorithm 5. Recall the regret is defined as

$$\text{Reg}(T) = \max_{i \in [K]} \sum_{t=1}^T \mu_i - \sum_{t=1}^T \mu_{J_t}, \quad (3.25)$$

where J_t is the node played by the algorithm (at epoch t), and $\mu_j = \mathbb{E}[\mathcal{L}(\mathcal{P}_{t,j})]$ is the expected hitting time of node j . The guarantees on the regret are stated below in Theorems 9 and 10.

Theorem 9. *Let T be any positive integer. Under Assumption 2, Algorithm 5 admits a regret of order*

$$\text{Reg}(T) = \tilde{\mathcal{O}} \left(\min \left\{ \frac{1}{(1-\rho)^2} \sqrt{\frac{T}{\alpha}}, \frac{1}{(1-\rho)^2} \sqrt{KT} \right\} \right),$$

where $\alpha = \min_{v \in V} \alpha_v$ (α_v defined in Definition 5).

Theorem 10. *On a problem instance that satisfies Assumption 2, Algorithm 5 achieves constant regret of order $\tilde{\mathcal{O}} \left(\sum_{v: \Delta_v > 0} \left(\Delta_v + \frac{1}{(1-\rho)^2 \Delta_v} \right) \right)$, where $\tilde{\mathcal{O}}$ omits absolute constants and logarithmic dependence on problem intrinsics.*

More details on Theorems 9 and 10 are in Appendices B.1.2 and B.1.3 for completeness. Also, the greedy algorithm also solves this problem. Discussions on the greedy algorithm are provided in Appendix B.1.4.

3.4 Adversarial Setting

In this section, we consider the case in which the network structure G_t changes over time, and study a version of this problem in which the adversary alters edge length across epochs: In each epoch, the adversary can arbitrarily pick edge lengths $l_{ij}^{(t)}$ from $[0, 1]$. Recall, in this case, the performance is measured by the regret against playing any fixed node $i \in [K]$:

$$\text{Reg}_i^{\text{adv}}(T) = \sum_{t=1}^T l_{t,i} - \sum_{t=1}^T l_{t,J_t},$$

where J_t is the node played in epoch t , $l_{t,j} = \mathbb{E}[\mathcal{L}(\mathcal{P}_{t,i})]$, and $\mathcal{L}(\mathcal{P}_{t,j})$ is defined in (3.1). Since $\mathcal{L}(\mathcal{P}_{t,j})$ concentrates around $l_{t,j}$, a high probability bound on $\text{Reg}_i^{\text{adv}}(T)$ naturally provides a high probability bound on $\sum_{t=1}^T \mathcal{L}(\mathcal{P}_{t,i}) - \sum_{t=1}^T \mathcal{L}(\mathcal{P}_{t,J_t})$.

We will use a new version of the exponential weight algorithm to solve this adversarial problem. Also, a high probability guarantee is provided using a new concentration lemma (Lemma 9). As background, exponential weights algorithms maintain a probability distribution over the choices. This probability distribution gives higher weights to historically more rewarding nodes. In each epoch, a node is sampled from this probability distribution, and information is recorded down. To symbolically describe the strategy, we first define some notations. We first extract a sample of $\mathcal{L}(\mathcal{P}_{t,j})$ from the trajectory \mathcal{P}_{t,J_t} , where J_t is the node played in epoch t .

Given the trajectory for epoch t $\mathcal{P}_{t,J_t} = (X_{t,0}, L_{t,1}, X_{t,1}, \dots, L_{t,H_t,J_t}, X_{t,H_t,J_t})$, we define, for $v \in [K]$,

$$Y_v(\mathcal{P}_{t,J_t}) = \max_{i:0 \leq i < H_t,J_t} \mathbb{I}_{[X_{t,i}=v]} \cdot \mathcal{L}_i(\mathcal{P}_{t,J_t}), \quad (3.26)$$

where $\mathcal{L}_i(\mathcal{P}_{t,J_t}) := \sum_{k=i+1}^{H_t,J_t} L_{t,k}$. In words, $\mathcal{L}_i(\mathcal{P}_{t,J_t})$ is the distance (sum of edge lengths) from the first occurrence of v to the absorbing node.

By the principle of Proposition 4, if node i is covered by trajectory \mathcal{P}_{t,J_t} , $Y_v(\mathcal{P}_{t,J_t})$ is a sample of $\mathcal{L}(\mathcal{P}_{t,i})$. We define, for the trajectory $\mathcal{P}_{t,J_t} = \{X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, \dots, L_{t,H_t,J_t}, X_{t,H_t,J_t}\}$,

$$Z_{t,v} := Y_v(\mathcal{P}_{t,J_t}), \quad \forall v \in [K],$$

where $Y_v(\mathcal{P}_{t,J_t})$ is defined above in (3.26).

Define $\mathbb{I}_{t,ij} := \mathbb{I}_{[i \in \mathcal{P}_{t,J_t} \text{ and } j \in \mathcal{P}_{t,J_t}, Y_i(\mathcal{P}_{t,J_t}) > Y_j(\mathcal{P}_{t,J_t})]}$. This indicator random variable is 1 iff i and j both show up in \mathcal{P}_{t,J_t} and the first occurrence of j is after the first occurrence of i . We then define

$$\widehat{q}_{t,ij} := \frac{\sum_{s=1}^{t-1} \mathbb{I}_{s,ij}}{N_t^+(i)},$$

which is an estimator of how likely j is visited via a trajectory starting at i . In other words, $\widehat{q}_{t,ij}$ is an estimator of $q_{ij} := \mathbb{P}(j \in \mathcal{P}_{t,i})$, which is the probability of j being visited by a trajectory from i .

Using the above defined $\widehat{q}_{t,ij}$ and sample $Z_{t,i}$, we define an estimator for $\frac{l_{t,j}-B}{B}$ as

$$\widehat{Z}_{t,i} := \frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} + \beta}{p_{ti} + \sum_{j \neq i} \widehat{q}_{t,ji} p_{tj}}, \quad \forall i \in [K], \quad (3.27)$$

where B, β are algorithm parameters ($\beta \leq \alpha$ and B to be specified later). Note that our estimator (3.27) is novel, and a small bias is introduced via β . In fact, a novel concentration result is derived for this estimator in Lemma 9. With the estimators $\widehat{Z}_{t,i}$, we define $\widehat{S}_{t,j} = \sum_{s=1}^{t-1} \widehat{Z}_{s,i}$. By convention, we set $\widehat{S}_{0,i} = 0$ for all $i \in [K]$. The probability of playing i in epoch t is defined as

$$p_{ti} := \begin{cases} \frac{1}{K}, & \text{if } t = 1, \\ \frac{\exp(\eta \widehat{S}_{t-1,i})}{\sum_{j=1}^K \exp(\eta \widehat{S}_{t-1,j})}, & \text{if } t \geq 2, \end{cases} \quad (3.28)$$

where η is the learning rate.

Against any arm $j \in [K]$, following the sampling rule (3.28) can guarantee an $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound. We now summarize our strategy in Algorithm 6, and state the performance guarantee in Theorem 11.

A high probability performance guarantee of Algorithm 6 is below in Theorem 11.

Theorem 11. *Let $\kappa := 1 + \sum_j \frac{1 - \sqrt{\alpha_j}}{1 + \sqrt{\alpha_j}}$. Fix any $i \in [K]$. If the time horizon T and algorithm parameters satisfies $\epsilon \leq \frac{1}{T}$, $B = \frac{\log \frac{(1-\rho)\epsilon}{KT}}{\log \rho}$, $\eta = \frac{1}{\sqrt{\kappa T}}$, $\beta = \frac{1}{\sqrt{\kappa T}} \leq \alpha$, and*

Algorithm 6

- 1: **Input:** A set of nodes $[K]$, transition matrix M , total number of epochs T , probability parameter $\epsilon \in \left(0, \frac{1}{(1-\rho)KT}\right)$. Algorithm parameters: $B = \frac{\log \frac{(1-\rho)\epsilon}{KT}}{\log \rho}$, $\eta = \frac{1}{\sqrt{\kappa T}}$, $\beta = \frac{1}{\sqrt{\kappa T}}$.
- 2: **for** $t = 1, 2, 3, \dots, T$ **do**
- 3: Randomly play node $J_t \in [K]$, such that

$$\mathbb{P}(J_t = i) = p_{ti}, \forall i \in [K]. \quad (3.29)$$

where p_{ti} is defined in (3.28).

- 4: Observe the trajectory \mathcal{P}_{t, J_t} . Update estimates $\hat{Z}_{t,j}$ according to (3.27).
-

$\epsilon \left(1 + KB + \frac{2K}{B(1-\rho)^3}\right) \leq \frac{3}{4}$, then with probability exceeding $1 - \tilde{\mathcal{O}}(\epsilon)$,

$$\text{Reg}_i^{\text{adv}}(T) \lesssim \frac{\log(1/\epsilon)}{\log(1/\rho)} \sqrt{\kappa T}, \quad (3.30)$$

where \lesssim omits constants, multiplicative terms that are (poly-)logarithmic in T and K , and additive terms of smaller order.

Note that $\sum_{j \in [K]} \frac{1 - \sqrt{\alpha_j}}{1 + \sqrt{\alpha_j}}$ can be orders-of-magnitude smaller than K , as we shown in the following example. In a complete symmetric graph where each node has probability of $1 - \rho$ of immediately hitting the absorbing node, $\alpha_i = \rho^2$ for all i . In this case, $\rho = \frac{1 - K^{-\lambda}}{1 + K^{-\lambda}}$ gives $\kappa := K^{1-\lambda}$ (for some $\lambda \in (0, 1)$). When ρ is close to 1, we know $\frac{1}{\log(1/\rho)} = \mathcal{O}\left(\frac{1}{1-\rho}\right)$. Thus in such cases the regret is of order $\tilde{\mathcal{O}}\left(\frac{1}{1-\rho} \sqrt{K^{1-\lambda} T}\right)$. In this example, the bound $\mathcal{O}\left(\frac{1}{1-\rho} \sqrt{K^{1-\lambda} T}\right)$ is significantly better than previous bounds on the standard adversarial MAB, online MDPs, and stochastic shortest path, since bounds for the latter are at best $\tilde{\mathcal{O}}\left(\frac{1}{1-\rho} \sqrt{KT}\right)$, where $\frac{1}{1-\rho}$ is effectively the reward range, or time horizon, or graph diameter.

In Figure 3.2, we provide a plot of $f(x) = \frac{1 - \sqrt{x}}{1 + \sqrt{x}}$ with $x \in [0, 1]$. This shows that the regret dependence on the connectivity of the graph is very nonlinear.

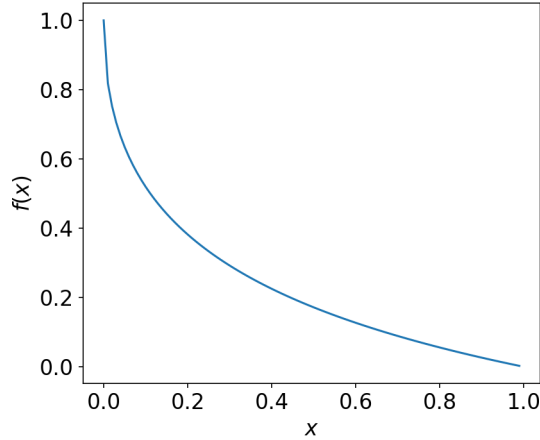


Figure 3.2: A plot of function $f(x) = \frac{1-\sqrt{x}}{1+\sqrt{x}}$, $x \in [0, 1]$. This shows that in Theorem 11, the dependence on graph connectivity is highly non-linear.

3.4.1 Analysis of Algorithm 6

In this section we provide proof for Theorem 11. To start with, we put forward the following notations for simplicity.

1. We write $\tilde{p}_{tj} = p_{tj} + \sum_{i \neq j} q_{ij} p_{ti}$, and $\hat{p}_{tj} = p_{tj} + \sum_{i \neq j} \hat{q}_{t,ij} p_{ti}$.
2. We use \mathcal{F}_t to denote the σ -algebra generated by all randomness up to the end of epoch t . We use $\mathcal{F}_{t,i}$ to denote the σ -algebra of all randomness up to the first occurrence of node i in epoch t (or end of epoch t if i is not visited in epoch t). We use \mathbb{E}_t to denote the expectation conditioning on \mathcal{F}_t , i.e., $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$.
3. Unless otherwise stated, we use \sum_t and \sum_j as shorthand for $\sum_{t=1}^T$ and $\sum_{j \in [K]}$, respectively.

The following two lemmas (Lemmas 6 and 7) depict properties of $\hat{q}_{t,ij}$, whose proof are in Appendix B.2.

Lemma 6. *For any t, i, j , it holds that $\mathbb{V}[\hat{q}_{t,ij}] = \tilde{\mathcal{O}}\left(\frac{1}{t}\right)$.*

Lemma 7. For any $\epsilon \in (0, 1)$, let

$$\mathcal{E}'_t := \left\{ |\hat{q}_{t,ij} - q_{ij}| \geq \sqrt{\frac{2\mathbb{V}[\hat{q}_{t,ij}] \log(T/\epsilon)}{N_{t-1}^+(i)}} + \frac{\log(T/\epsilon)}{3N_{t-1}^+(i)}, N_t^+(i) \geq \alpha t - \sqrt{t \log(TK/\epsilon)}, \forall i, j \in [K] \right\}.$$

It holds that $\mathbb{P}(\mathcal{E}'_t) \geq 1 - \frac{2\epsilon}{T}$ and under \mathcal{E}'_t ,

$$\hat{q}_{t,ij} = q_{ij} \pm \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right), \quad (3.31)$$

where $\alpha = \min_{i \in [K]} \alpha_j$.

The following lemma provides properties of the random variable $Z_{t,i}$.

Lemma 8. For any B , let $\mathcal{E}_T(B) := \{Z_{t,j} \leq B \text{ for all } t = 1, 2, \dots, T, \text{ and } j \in [K]\}$. For any $\epsilon \in (0, 1)$ and $B = \frac{\log\left(\frac{(1-\rho)\epsilon}{KT}\right)}{\log \rho}$, it holds that

$$\mathbb{P}(\mathcal{E}_T(B)) \geq 1 - \epsilon, \quad \mathbb{E}[Z_{t,i} | \text{not } \mathcal{E}_T(B)]$$

and

$$\leq KB + \frac{K}{(1-\rho)^2},$$

and

$$\mathbb{E}[Z_{t,i}^2 | \text{not } \mathcal{E}_T(B)] \leq KB^2 + \frac{2K}{(1-\rho)^3}.$$

The first equation Lemma 8 is a high probability event, and the last two equations in are a type of memorylessness property. The proof of Lemma 8 can be found in Appendix B.2.

Next, we provide a novel concentration result in Lemma 9. This lemma provides a bound for our reward estimator (3.27), and gives insights for adversarial bandit problems where the adversary stochastically picks rewards by specifying reward distributions.

Lemma 9. For any $\epsilon \in (0, 1)$ and $T \in \mathbb{N}$, such that $\epsilon \leq \frac{1}{T}$ and $T \geq 10$,

$$\mathbb{P}\left(\sum_t \frac{l_{t,i} - B}{B} - \sum_t \hat{Z}_{t,i} \leq \frac{\log(T/\epsilon^2)}{\beta}\right) \geq 1 - \tilde{\mathcal{O}}(\epsilon).$$

Proof. By a total law of expectation, we have

$$\mathbb{E}_{t-1} \left[Z_{t,i} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} \right] = \mathbb{E}_{t-1} \left[\mathbb{E} \left[Z_{t,i} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} \middle| \mathcal{F}_{t,i} \right] \right] = \mathbb{E}_{t-1} \left[l_{t,i} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} \right] = l_{t,i} \tilde{p}_{ti}. \quad (3.32)$$

Pick ϵ so that $\epsilon \left(1 + KB + \frac{2K}{B(1-\rho)^3} \right) \leq \frac{3}{4}$ (this is doable since $B = \mathcal{O}(\log(1/\epsilon))$). By Lemma 8, we have

$$\begin{aligned} \mathbb{E}_{t-1} \left[\left(\frac{Z_{t,i} - B}{B} \right)^2 \right] &\leq \mathbb{E}_{t-1} \left[\left(\frac{Z_{t,i} - B}{B} \right)^2 \middle| \mathcal{E}_T(B) \right] \\ &\quad + \mathbb{E}_{t-1} \left[\left(\frac{Z_{t,i} - B}{B} \right)^2 \middle| \text{not } \mathcal{E}_T(B) \right] (1 - \mathbb{P}(\mathcal{E}_T(B))) \\ &\leq \frac{1}{4} + \epsilon \left(1 + KB + \frac{2K}{B(1-\rho)^3} \right) \leq 1. \end{aligned} \quad (3.33)$$

By Lemma 7 and a Taylor expansion, we know, under event $\cap_{t=1}^T \mathcal{E}'_t$,

$$\frac{1}{\widehat{p}_{tj}} = \frac{1}{\tilde{p}_{tj}} \pm \tilde{\mathcal{O}} \left(\frac{\log(1/\epsilon)}{t} \right), \quad \forall t \in [T], j \in [K] \quad (3.34)$$

We can use (3.34) to get

$$\begin{aligned} &\mathbb{E}_{t-1} \left[\exp \left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{Z_{t,i} - B \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} + \beta}{p_{ti} + \sum_{j \neq i} \widehat{q}_{t,j} p_{tj}} \right) \right) \middle| \mathcal{E}_T(B) \cap \left(\cap_{t=1}^T \mathcal{E}'_t \right) \right] \\ &= \mathbb{E}_{t-1} \left[\exp \left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{Z_{t,i} - B \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} + \beta}{\tilde{p}_{ti}} \right) + \tilde{\mathcal{O}} \left(\frac{\log(1/\epsilon)}{t} \right) \right) \middle| \mathcal{E}_T(B) \cap \left(\cap_{t=1}^T \mathcal{E}'_t \right) \right] \\ &= \exp \left(-\frac{\beta^2}{\tilde{p}_{ti}} + \tilde{\mathcal{O}} \left(\frac{\log(1/\epsilon)}{t} \right) \right). \\ &\mathbb{E}_{t-1} \left[\exp \left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{Z_{t,i} - B \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} + \beta}{\tilde{p}_{ti}} \right) \right) \middle| \mathcal{E}_T(B) \cap \left(\cap_{t=1}^T \mathcal{E}'_t \right) \right]. \end{aligned} \quad (3.35)$$

Under event $\mathcal{E}_T(B)$, we have $-\beta \left(\frac{Z_{t,i} - B \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} + \beta}{\tilde{p}_{ti}} \right) \leq 1$, since $\tilde{p}_{ti} \geq \alpha$ and $\beta \leq \alpha$ (and

$B \geq \frac{1}{1-\rho} \geq l_{t,i}$. Since $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$, we have

$$\begin{aligned}
& \mathbb{E}_{t-1} \left[\exp \left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]}}{\tilde{p}_{ti}} \right) \right) \middle| \mathcal{E}_T(B) \cap \left(\bigcap_{t=1}^T \mathcal{E}'_t \right) \right] \\
& \leq 1 + \mathbb{E}_{t-1} \left[\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]}}{\tilde{p}_{ti}} \right) \middle| \mathcal{E}_T(B) \cap \left(\bigcap_{t=1}^T \mathcal{E}'_t \right) \right] \\
& \quad + \mathbb{E}_{t-1} \left[\left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]}}{\tilde{p}_{ti}} \right) \right)^2 \middle| \mathcal{E}_T(B) \cap \left(\bigcap_{t=1}^T \mathcal{E}'_t \right) \right] \\
& \leq \frac{1}{1-3\epsilon} \left(1 + \mathbb{E}_{t-1} \left[\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]}}{\tilde{p}_{ti}} \right) \right] \right. \\
& \quad \left. + \mathbb{E}_{t-1} \left[\left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]}}{\tilde{p}_{ti}} \right) \right)^2 \right] \right) \tag{3.36}
\end{aligned}$$

where (3.36) uses that $\mathbb{E}[A | \mathcal{E}_T(B) \cap (\bigcap_{t=1}^T \mathcal{E}'_t)] \leq \frac{\mathbb{E}[A]}{\mathbb{P}(\mathcal{E}_T(B) \cap (\bigcap_{t=1}^T \mathcal{E}'_t))}$ for any event A , and $\mathbb{P}(\mathcal{E}_T(B) \cap (\bigcap_{t=1}^T \mathcal{E}'_t)) \geq 1 - 3\epsilon$.

Since

$$\begin{aligned}
\mathbb{E}_{t-1} \left[\mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} \right] &= \tilde{p}_{tj} \\
\mathbb{E}_{t-1} \left[(Z_{t,i} - B) \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} \right] &= (l_{t,j} - B) \tilde{p}_{tj} \\
\mathbb{E}_{t-1} \left[\left(\frac{Z_{t,i} - B}{B} \right)^2 \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]} \right] &\leq \tilde{p}_{tj},
\end{aligned}$$

we have

$$\begin{aligned}
& 1 + \mathbb{E}_{t-1} \left[\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]}}{\tilde{p}_{ti}} \right) \right] \\
& \quad + \mathbb{E}_{t-1} \left[\left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}]}}{\tilde{p}_{ti}} \right) \right)^2 \right] \\
& = 1 - \frac{\beta^2}{B^2} (l_{t,i} - B)^2 + \frac{\beta^2}{\tilde{p}_{ti}} \leq \exp \left(\frac{\beta^2}{\tilde{p}_{ti}} \right), \tag{3.37}
\end{aligned}$$

where the last inequality uses $1 + x \leq \exp(x)$.

We can now combine (3.35), (3.36) and (3.37) and get

$$\begin{aligned} & \mathbb{E}_{t-1} \left[\exp \left(\frac{\beta}{B} (l_{t,i} - B) - \beta \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}] + \beta}}{p_{ti} + \sum_{j \neq i} \widehat{q}_{t,j} p_{tj}} \right) \right) \middle| \mathcal{E}_T(B) \cap \left(\cap_{t=1}^T \mathcal{E}'_t \right) \right] \\ & \leq \frac{1}{1-3\epsilon} \exp \left(\widetilde{\mathcal{O}} \left(\frac{\log(1/\epsilon)}{t} \right) \right) \end{aligned}$$

Let $X = \frac{\beta}{B} \sum_{t=1}^T (l_{t,i} - B) - \beta \sum_{t=1}^T \left(\frac{\frac{Z_{t,i}-B}{B} \mathbb{I}_{[i \in \mathcal{P}_{t,J_t}] + \beta}}{p_{ti} + \sum_{j \neq i} \widehat{q}_{t,j} p_{tj}} \right)$ for simplicity. Note that

$$X = \beta \left(\sum_{t=1}^T \frac{l_{t,i} - B}{B} - \sum_{t=1}^T \widehat{Z}_{t,i} \right).$$

We combine (3.35), (3.36) and (3.37) and get

$$\mathbb{E} \left[e^X \middle| \mathcal{E}_T(B) \cap \left(\cap_{t=1}^T \mathcal{E}'_t \right) \right] \leq \prod_{t=1}^T \exp \left(\log \frac{1}{1-3\epsilon} + \widetilde{\mathcal{O}} \left(\frac{\log(1/\epsilon)}{t} \right) \right) = \left(\frac{1}{1-3\epsilon} \right)^T \widetilde{\mathcal{O}}(T/\epsilon).$$

By (conditional) Markov inequality and (3.37), we have

$$\mathbb{P} \left(\frac{X}{\beta} \geq \frac{\log(T/\epsilon^2)}{\beta} \middle| \mathcal{E}_T(B) \cap \left(\cap_{t=1}^T \mathcal{E}'_t \right) \right) \leq \frac{\epsilon^2}{T} \mathbb{E} e^X \leq \left(\frac{1}{1-3\epsilon} \right)^T \widetilde{\mathcal{O}}(\epsilon).$$

Since $\left(\frac{1}{1-3\epsilon} \right)^T \leq 40$ when $\epsilon \leq \frac{1}{T}$ and $T \geq 10$, we can apply a union bound (to remove the conditioning on $\mathcal{E}_T(B) \cap \left(\cap_{t=1}^T \mathcal{E}'_t \right)$) and conclude the proof. □

A final ingredient of proving Theorem 11 is Lemma 10, which is a high probability concentration result. The proof of Lemma 10 is in Appendix B.2.

Lemma 10. *With probability at least $1 - 6\epsilon$, we have*

$$\begin{aligned} & \sum_{t=1}^T \sum_j \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} \leq \kappa T + \widetilde{\mathcal{O}} \left(\sqrt{T \log(1/\epsilon)} \right), \\ & \sum_t \sum_j p_{tj} \widehat{Z}_{t,j} - \sum_t \frac{l_{t,J_t} - B}{B} \leq \kappa \beta T + \mathcal{O} \left(\sqrt{(1 + \beta \kappa) T \log(1/\epsilon)} \right). \end{aligned}$$

With the above preparation, we can now prove Theorem 11.

Proof of Theorem 11. By the exponential weights argument [LW94, ACBFS02], it holds that, under event $\mathcal{E}_T(B)$,

$$\sum_{t=1}^T \widehat{Z}_{t,i} - \sum_{t=1}^T \sum_j p_{tj} \widehat{Z}_{t,j} \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^T \sum_j \frac{p_{tj}}{\widehat{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]}$$

We use the first equation in Lemma 10 and a union bound (to remove the dependence on $\mathcal{E}_T(B)$) to get, with probability at least $1 - \widetilde{\mathcal{O}}(\epsilon)$,

$$\sum_{t=1}^T \widehat{Z}_{t,i} - \sum_{t=1}^T \sum_j p_{tj} \widehat{Z}_{t,j} \leq \frac{\log K}{\eta} + \eta \kappa T + \widetilde{\mathcal{O}}\left(\eta \sqrt{T \log(1/\epsilon)}\right). \quad (3.38)$$

We now combine above results to get, with probability at least $1 - \mathcal{O}(\epsilon)$,

$$\begin{aligned} \sum_t l_{t,i} - \sum_t l_{t,J_t} &= B \left(\sum_t \frac{l_{t,i} - B}{B} - \sum_t \frac{l_{t,J_t} - B}{B} \right) \\ &\stackrel{\textcircled{1}}{\leq} B \left(\sum_t \widehat{Z}_{t,i} - \sum_t \sum_j p_{tj} \widehat{Z}_{t,i} + \frac{\log(T/\epsilon^2)}{\beta} \right. \\ &\quad \left. + \kappa \beta T + \mathcal{O}\left(\sqrt{(1 + \beta \kappa) T \log(1/\epsilon)}\right) \right) \\ &\stackrel{\textcircled{2}}{\leq} B \left(\frac{\log K}{\eta} + \eta \kappa T + \frac{\log(T/\epsilon^2)}{\beta} + \kappa \beta T \right. \\ &\quad \left. + \mathcal{O}\left(\eta \sqrt{T \log(1/\epsilon)}\right) + \mathcal{O}\left(\sqrt{(1 + \beta \kappa) T \log(1/\epsilon)}\right) \right), \end{aligned}$$

where $\textcircled{1}$ uses Lemma 9 and the second inequality in Lemma 10, and $\textcircled{2}$ uses (3.38).

Setting $\eta = \sqrt{\frac{1}{\kappa T}}$ and $\beta = \sqrt{\frac{1}{\kappa T}}$ concludes the proof. □

3.4.2 Lower Bound for the Adversarial Setting

We also provide a lower bound for the adversarial case, which is summarized in Theorem 12. To prove this theorem, we first tweak the problem so that the adversary chooses distribution over edge lengths (Proposition 9 in Appendix B.2.2), and give a bound under this randomization. With this tweak, the sample space of a single trajectory is $\cup_{h=1}^{\infty} ([0, 1] \cup [K])^h$, which is much larger than $[0, 1]$. This means a trajectory carries much more information

than a single reward sample. We prove that the lower bound is of order $\Omega(\sqrt{T})$, even though much more information is available.

Theorem 12. Fix any $T > \sqrt{128 \log 8}$ and $\sigma < \frac{1}{7}$. On a set of K transient nodes, there exists a sequence of edge lengths and a node $i \in [K]$ and a transition probability matrix, such that for any policy, the regret incurred by any π against i satisfies

$$\mathbb{P}_{\mathfrak{J}, \pi} \left(\text{Reg}_j^{\text{adv}}(T) \geq \min \left\{ \sqrt{\frac{(1 - Kp)^2 \sigma^2 T}{32p}}, \sqrt{\frac{(K - 1)(1 - Kp) \sigma^2 T}{32 \left(1 + \frac{p}{1 - Kp}\right)}} \right\} \right) \geq \frac{1}{8}.$$

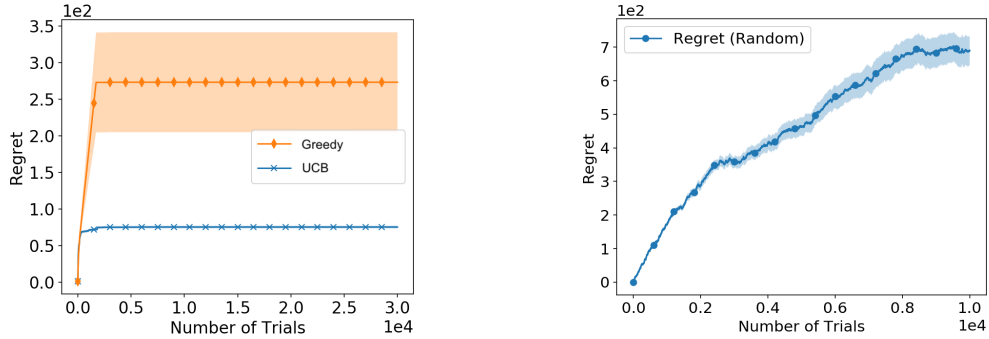


Figure 3.3: Left: Results for stochastic setting. The line labelled UCB corresponds to Algorithm 5 and the line labelled Greedy corresponds to Algorithm 7. Right: Results for the adversarial setting. Each line is averaged over 10 runs. The shaded areas (around the solid lines) indicate one standard deviation below and above the average. For the adversarial case, the regret is measured against arm 0, which, by construction, has largest hitting time in expectation.

Similar to the proof for Theorem 8, we construct two problem instances \mathfrak{J} and \mathfrak{J}' such that no policy can quickly tell the difference between them. We again use $\mathbb{P}_{\mathfrak{J}, \pi}$ (resp. $\mathbb{P}_{\mathfrak{J}', \pi}$) to denote the probability measure generated by playing π on \mathfrak{J} (resp. \mathfrak{J}'). Similar to the proof for Theorem 8, the sample space (on which both $\mathbb{P}_{\mathfrak{J}, \pi}$ and $\mathbb{P}_{\mathfrak{J}', \pi}$ is defined) is different from a regular bandit problem. If we execute π for T epochs, the sample space is then $\left(\cup_{h=1}^{\infty} ([0, 1] \times [K])^h \right)^T$ (with the σ -algebra generated by singletons in $[K]$ and Borel

sets in $[0, 1]$). This is because (1) each trajectory can be arbitrarily long, (2) the nodes on the trajectory can be any of $[K]$, and (3) each edge on the trajectory can take values from $[0, 1]$. Specifically, for each trajectory $\mathcal{P}_{t, J_t} = (X_{t,0}, L_{t,1}, X_{t,1}, \dots, L_{t, H_{t, J_t}}, X_{t, H_{t, J_t}})$, H_{t, J_t} can be any positive integer, $X_{t,i}$ can be any integer from $[K]$, and each $L_{t,i}$ can be any number from $[0, 1]$. This sample space is fundamentally different from the space generated by interaction with a standard K -armed bandit problem for T rounds, which is $[0, 1]^{KT}$. We use the Markov property of random walks to handle this difficulty. Specifically, for any fixed i and j , conditioning on $X_{t,i} = j$ being known, the space generated by $L_{t,i+1}, X_{t,i+1}, L_{t,i+2}, X_{t,i+2}, \dots$ is identical to the space generated by $L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \dots$ conditioning on $X_{t,0} = j$ being fixed. A proof of Theorem 12 can be found in Appendix B.2.1.

3.5 Experiments

We deploy our algorithms on a problem with 9 transient nodes. For stochastic setting (left subfigure in Figure 3.3), we have $l_{ij} = 1$ for $i \in [9]$ and $j \in [9] \cup \{*\}$. The transition probabilities among transient nodes are

$$m_{ij} = \begin{cases} 0.6, & \text{if } i = j = 1, \\ 0.4, & \text{if } i = j \text{ and } i \neq 1, \\ 0.1, & \text{if } i = j \pm 1 \pmod{9}, \\ 0, & \text{otherwise.} \end{cases}$$

For adversarial setting (right subfigure in Figure 3.3), the transition probabilities among transient nodes are

$$m_{ij} = \begin{cases} 0.3, & \text{if } i = j, \\ 0.1, & \text{if } i = j \pm 1 \pmod{9}, \\ 0, & \text{otherwise.} \end{cases}$$

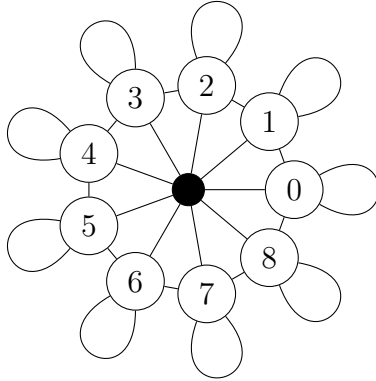


Figure 3.4: The network structure for experiments. The dark node at the center is the absorbing node $*$, and nodes labelled with numbers are transient nodes. Nodes without edges connecting them visits each other with zero probability.

The edge lengths are sampled from Gaussian distributions and truncated to between 0 and 1. Specifically, for all $t = 1, 2, \dots, T$,

$$l_{ij}^{(t)} = \begin{cases} \text{clip}_{[0,1]}(W_t + 0.5), & \text{if } i = 0 \text{ and } j = * \\ \text{clip}_{[0,1]}(W_t), & \text{if } i \neq 0 \text{ and } j = * \\ 1, & \text{otherwise,} \end{cases}$$

where $\text{clip}_{[0,1]}(z)$ takes a number z and clips it to $[0, 1]$, and $W_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0.5, 0.1)$. The results (Figure 3.3) empirically consolidate our theorems:

3.6 Conclusion

In this paper, we propose the problem **(P)** motivated by online advertisement, and study it from a bandit learning perspective. We study both the stochastic setting and the adversarial setting for this problem. Our paper provides a comprehensive study of this important problem **(P)** from a bandit learning perspective.

Chapter 4

Conclusion

My thesis focuses on multi-armed bandit problem with new feedback structures. In Chapter 2, multi-armed bandit problems in metric spaces are studied. In particular, bandits for BMO functions are introduced and examined. In Chapter 3, multi-armed bandit problems with random walk trajectories as feedback are studied. Algorithms leveraging such feedback structures are introduced, and corresponding lower bounds are also derived. Both chapters are based on previously released materials, as listed below:

1. Tianyu Wang, Weicheng Ye, Dawei Geng, Cynthia Rudin, Towards Practical Lipschitz Bandits, *ACM-IMS Foundations of Data Science (FODS)*, 2020.
2. Tianyu Wang and Cynthia Rudin, Bandits for BMO Functions, *International Conference on Machine Learning (ICML)*, 2020.
3. Tianyu Wang, Lin F. Yang, Zizhuo Wang, Towards Fundamental Limits of Multi-armed Bandits with Random Walk Feedback, *arXiv:2011.01445*.

Other Works

Apart from aforementioned directions that focuses on multi-armed bandit with new feedback structures, I have also worked on other problems, as listed below in chronological order:

- Interpretable and scalable matching methods for causal inference [WMA⁺17]. We use machine learning algorithms to learn which features to match on. This effectively learns a distance metric in the feature space, while maintaining interpretability of the outputs. Our implementations use bit-vector manipulation and database techniques, which gives a level of scalability that was not achieved by previous methods.

- Adaptive model-based (finite-horizon) reinforcement learning algorithms in metric spaces [SWJ⁺20]. We introduce the technique of adaptive discretization to design an efficient model-based episodic reinforcement learning algorithm in large (potentially continuous) state-action spaces. Our algorithm is based on optimistic one-step value iteration extended to maintain an adaptive discretization of the space. Our bounds are obtained via a modular proof technique which can potentially extend to incorporate additional structure on the problem.
- A (finite-horizon) Linear Quadratic Regulator problem with a new low-rank transition structure [WY20]. We propose an algorithm that utilizes the intrinsic system low-rank structure for efficient learning. For problems of rank- m , our algorithm achieves a K -episode regret bound of order $\tilde{O}(m^{3/2}K^{1/2})$. Consequently, the sample complexity of our algorithm only depends on the rank, m , rather than the ambient dimension, d , which can be orders-of-magnitude larger.

Appendices

Appendix A

Supplementary Materials for Chapter 2

For readability, we reiterate the lemma statements before presenting the proofs.

A.1 Proof of Lemma 2

Lemma 2 . *Let f be the reward function. For any f -admissible $\delta \geq 0$, let $S^\delta := \{a \in [0, 1]^d : f(a) > f^\delta\}$. Then we have S^δ measurable and $\mu(S^\delta) = \delta$.*

Proof. Recall

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

We consider the following two cases.

Case 1: $f^\delta \in F^\delta$, then by definition (of F^δ), $\mu(S^\delta) = \delta$.

Case 2: $f^\delta \notin F^\delta$ (F^δ is left open). Then by definition of the infimum operation, for any $i = 1, 2, 3, \dots$, there exists $z_i \in F^\delta$, such that $f^\delta < z_i \leq f^\delta + \frac{1}{i}$. Thus $\lim_{i \rightarrow \infty} z_i = f^\delta$.

We know that f is Lebesgue measurable, since

$$f \in BMO(\mathbb{R}^d, \mu) \Rightarrow f \text{ is Lebesgue measurable.}$$

Let us define $S_i := \{a \in [0, 1]^d : f(a) > z_i\}$. By this definition, $S_1 \subseteq S_2 \subseteq S_3 \dots$. Also S_i is Lebesgue measurable, since it is the pre-image of the open set (z_i, ∞) under the Lebesgue measurable function f . By the above construction of S_i , we have $\mu(S_i) = \delta$ for all $i = 1, 2, 3, \dots$. By continuity of measure from below,

$$\mu(\cup_{i=1}^{\infty} S_i) = \lim_{i \rightarrow \infty} \mu(S_i). \tag{A.1}$$

We also have $S^\delta = \cup_{i=1}^{\infty} S_i$. This is because

- (1) $S^\delta \supseteq \cup_{i=1}^\infty S_i$, since by definition, $S_i \subseteq S^\delta$ for all $i = 1, 2, 3, \dots$;
- (2) $S^\delta \subseteq \cup_{i=1}^\infty S_i$, since $\lim_{i \rightarrow \infty} z_i = f^\delta$ and therefore every element in S^δ is an element in $\cup_{i=1}^\infty S_i$.

Hence,

$$\mu(S^\delta) = \mu(\cup_{i=1}^\infty S_i) = \lim_{i \rightarrow \infty} \mu(S_i) = \delta, \quad (\text{A.2})$$

where the last equality uses $\mu(S_i) = \delta$ for all $i = 1, 2, 3, \dots$. \square

A.2 Proof of Lemma 3

First we state a corollary of the Azuma's inequality. This lemma is Proposition 34 by Tao and Vu [TV15], and can be derived using Lemma 3.1 by Vu [Vu02]. We prove a proof below for completeness. We will use this lemma to prove Lemma 3.

Lemma 11 (Proposition 34 in [TV15]). *Consider a martingale sequence X_1, X_2, \dots adapted to filtration $\mathcal{F}_1, \mathcal{F}_2, \dots$. For constants $c_1, c_2, \dots < \infty$, we have*

$$\mathbb{P} \left(|X_n - X_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2} \right) \leq 2 \exp \left(-\frac{\lambda^2}{2} \right) + \sum_{i=1}^n \mathbb{P}(|X_i - X_{i-1}| > c_i). \quad (\text{A.3})$$

Proof. Define the “good event” $\mathcal{G}_n := \{|X_i - X_{i-1}| \leq c_i, \text{ for all } i \leq n\}$. Rewrite the above probability as

$$\begin{aligned} & \mathbb{P} \left(|X_n - X_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2} \right) \\ &= \mathbb{P} \left(|X_n - X_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2} \mid \mathcal{G}_n \right) \mathbb{P}(\mathcal{G}_n) + \mathbb{P} \left(|X_n - X_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2} \mid \bar{\mathcal{G}}_n \right) (1 - \mathbb{P}(\mathcal{G}_n)) \\ &\leq \mathbb{P} \left(|X_n - X_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2} \mid \mathcal{G}_n \right) + (1 - \mathbb{P}(\mathcal{G}_n)). \end{aligned} \quad (\text{A.4})$$

In (A.4), the first term can be bounded by applying Azuma's inequality for martingales of bounded difference, and the second term is the probability of there existing at least one

difference being large. For the first term, we define $X'_i := X_i \mathbb{I}[|X_i - X_{i-1}| \leq c_i]$. It is clear that $\{X'_i\}_i$ is also martingale sequence adapted to $\mathcal{F}_1, \mathcal{F}_2, \dots$. Using this new sequence, we have

$$\mathbb{P} \left(|X_n - X_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2} \middle| \mathcal{G}_n \right) = \mathbb{P} \left(|X'_n - X'_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2} \right) \leq 2 \exp \left(-\frac{\lambda^2}{2} \right),$$

where the last inequality is a direct consequence of Azuma's inequality.

Finally, we take a union bound and a complement to get

$$\mathbb{P}(\overline{\mathcal{G}}_t) \leq \sum_{i=1}^n \mathbb{P}(|X_i - X_{i-1}| > c_i).$$

This finishes the proof. □

Lemma 3. *Pick $T \geq 1$ and $\epsilon \in (0, 1)$. 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 , where*

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

$$H_t(q) = \frac{(\Psi + D_E) \sqrt{2 \log(2T^2/\epsilon)}}{\sqrt{\tilde{n}_t(q)}}.$$

Proof. Case I: We first take care of the case when q contains at least one observation.

Define

$$\mathcal{F}'_i := \sigma(Q_1, A_1, Y_1, \dots, Q_{i-1}, A_{i-1}, Y_{i-1}, Q_i, A_i).$$

By our partition refinement rule, we have that for any t, t' such that $t \geq t'$ and $q \in \mathcal{Q}_t$, there exists $q' \in \mathcal{Q}_{t'}$ such that $q \subseteq q'$. Thus for any $i \leq t$, and any $q \in \mathcal{Q}_t$, we have either $Q_i \supseteq q$ or $Q_i \cap q = \emptyset$ (Q_i is the cube played at time $i \leq t$). Thus, we have

$$\begin{aligned} \mathbb{E} [Y_i \mathbb{I}_{[A_i \in q]} | \mathcal{F}'_i] &= \begin{cases} \langle f \rangle_q \mathbb{I}_{[A_i \in q]}, & \text{if } Q_i \supseteq q, \\ 0, & \text{if } Q_i \cap q = \emptyset, \end{cases} \\ &= \langle f \rangle_q \mathbb{I}_{[A_i \in q]}, \end{aligned} \tag{A.5}$$

where $\mathbb{I}_{[A_i \in q]}$ is \mathcal{F}'_i -measurable. In (A.5), the two cases are exhaustive as discussed above.

Therefore the sequence $\left\{ \left(Y_i - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right\}_i$ is a (skipped) martingale difference sequence adapted to \mathcal{F}'_i , with the skipping event $\mathbb{I}_{[A_i \in q]}$ being \mathcal{F}'_i -measurable.

Let A' be a uniform random variable drawn from the cube q . We have

$$\begin{aligned}
\mathbb{P} \left(\left| \left(f(A_i) - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| > \Psi \right) &= \begin{cases} \mathbb{P} \left(\left| f(A') - \langle f \rangle_q \right| > \Psi \right), & \text{if } A_i \in q, \\ 0, & \text{otherwise .} \end{cases} \\
&\leq \mathbb{P} \left(\left| f(A') - \langle f \rangle_q \right| > \Psi \right) \\
&= \frac{\mu \left(a \in q : \left| \left(f(a) - \langle f \rangle_q \right) \right| > \Psi \right)}{\mu(q)} \\
&\leq \frac{\mu(q) \exp(-\Psi)}{\mu(q)} \leq \frac{\epsilon}{T^2}, \tag{A.6}
\end{aligned}$$

where (A.6) is from the John-Nirenberg inequality.

Next, since

$$\left| \left(Y_i - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| \leq \left| \left(f(A_i) - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| + \left| \left(Y_i - f(A_i) \right) \mathbb{I}_{[A_i \in q]} \right|,$$

we have

$$\begin{aligned}
&\mathbb{P} \left(\left| \left(Y_i - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| \leq \Psi + D_{\mathcal{E}} \right) \tag{A.7} \\
&\geq \mathbb{P} \left(\left| \left(f(A_i) - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| + \left| \left(Y_i - f(A_i) \right) \mathbb{I}_{[A_i \in q]} \right| \leq \Psi + D_{\mathcal{E}} \right) \\
&\geq \mathbb{P} \left(\left| \left(f(A_i) - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| \leq \Psi \quad \text{and} \quad \left| \left(Y_i - f(A_i) \right) \mathbb{I}_{[A_i \in q]} \right| \leq D_{\mathcal{E}} \right) \\
&= 1 - \mathbb{P} \left(\left| \left(f(A_i) - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| > \Psi \quad \text{or} \quad \left| \left(Y_i - f(A_i) \right) \mathbb{I}_{[A_i \in q]} \right| > D_{\mathcal{E}} \right) \\
&\geq 1 - \mathbb{P} \left(\left| \left(f(A_i) - \langle f \rangle_q \right) \mathbb{I}_{[A_i \in q]} \right| > \Psi \right) - \mathbb{P} \left(\left| \left(Y_i - f(A_i) \right) \mathbb{I}_{[A_i \in q]} \right| > D_{\mathcal{E}} \right), \tag{A.8}
\end{aligned}$$

where (A.8) uses a union bound.

By a union bound and the John-Nirenberg inequality, for any $i \leq t$, and $q \in \mathcal{Q}_t$, we

have

$$\begin{aligned} & \mathbb{P}\left(\left|(Y_i - \langle f \rangle_q) \mathbb{I}_{[A_i \in q]}\right| > \Psi + D_{\mathcal{E}}\right) \\ &= 1 - \mathbb{P}\left(\left|(Y_i - \langle f \rangle_q) \mathbb{I}_{[A_i \in q]}\right| \leq \Psi + D_{\mathcal{E}}\right) \\ &\leq \mathbb{P}\left(\left|(f(A_i) - \langle f \rangle_q) \mathbb{I}_{[A_i \in q]}\right| > \Psi\right) + \mathbb{P}\left(\left|(f(A_i) - Y_i) \mathbb{I}_{[A_i \in q]}\right| > D_{\mathcal{E}}\right) \end{aligned} \quad (\text{A.9})$$

$$= \mathbb{P}\left(\left|(f(A_i) - \langle f \rangle_q) \mathbb{I}_{[A_i \in q]}\right| > \Psi\right) \quad (\text{A.10})$$

$$\leq \frac{\epsilon}{T^2}, \quad (\text{A.11})$$

where (A.9) uses (A.8), (A.10) uses the boundedness of noise **(N1)**, and (A.11) uses (A.6).

To put it all together, we can apply Lemma 11 to the (skipped) martingale

$\left\{\sum_{j=1}^i (Y_j - \langle f \rangle_q) \mathbb{I}_{[A_j \in q]}\right\}_{i=1,2,\dots}$ (with $c_i = \Psi + D_{\mathcal{E}}$, $\lambda = \sqrt{2 \log(2T^2/\epsilon)}$, and $X_i = (Y_i - \langle f \rangle_q) \mathbb{I}_{[A_i \in q]}$) to get for $T \geq 2$ and a cube $q \in \mathcal{Q}_t$ such that $n_t(q) > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{t-1} (Y_i - \langle f \rangle_q) \mathbb{I}_{[A_i \in q]}\right| > (\Psi + D_{\mathcal{E}}) \sqrt{n_t(q)} \sqrt{2 \log(2T^2/\epsilon)}\right) \quad (\text{A.12})$$

$$\leq 2 \exp\left(-\frac{2 \log(2T^2/\epsilon)}{2}\right) + \sum_{i=1}^{t-1} \mathbb{P}\left(\left|(Y_i - \langle f \rangle_q) \mathbb{I}_{[A_i \in q]}\right| > \Psi + D_{\mathcal{E}}\right) \quad (\text{A.13})$$

$$\leq \frac{\epsilon}{T^2} + (t-1) \frac{\epsilon}{T^2} \leq \frac{\epsilon}{T},$$

where (A.12) uses Lemma 11, (A.13) uses (A.11) for the summation term.

Since $n_t(q) > 0$, we use

$$m_t(q) = \frac{1}{n_t(q)} \sum_{i=1}^{t-1} Y_i \mathbb{I}_{[A_i \in q]},$$

to rewrite (A.12) by dividing both sides by $n_t(q)$ to get

$$\mathbb{P}\left(\left|m_t(q) - \langle f \rangle_q\right| > \frac{(\Psi + D_{\mathcal{E}})}{\sqrt{n_t(q)}} \sqrt{2 \log(2T^2/\epsilon)}\right) \leq \frac{\epsilon}{T}.$$

Case II: Next, we consider the case where q contains no observations.

In order to do this, we need Propositions 5 and 6, which are proved in A.2.1 and A.2.2.

Proposition 5. *Following (2.50), the minimal cube measure is at least η . Thus the maximal number of cubes produced by Algorithm 3 is $\frac{1}{\eta}$, since the arm space is of measure 1.*

Proposition 6. For a function $f \in BMO(\mathbb{R}^d, \mu)$, and rectangles q_0, q_1, \dots, q_k such that $q_0 \subseteq q_1 \subseteq q_2 \subseteq \dots \subseteq q_k$, and constant $K \geq 1$ such that $K\mu(q_i) \geq \mu(q_{i+1})$ for all $i \in [0, k-1]$, we have

$$\left| \langle f \rangle_{q_0} - \langle f \rangle_{q_k} \right| \leq Kk \|f\|.$$

Let's continue with the proof of Lemma 3. By the lower bound on cube measure (Proposition 5), we know that $\mu(q) \geq \eta$ for any q generated by the algorithm. Let us construct a sequence of hyper-rectangles $q = q_0, q_1, \dots, q_k \subseteq [0, 1]^d$, such that $q_i \subseteq q_{i+1}$ for $i = 0, 1, \dots, k$, $\mu(q_{i+1}) = 2\mu(q_i)$, and $q_k = [0, 1]^d$. Since q is generated by the algorithm, we know $\mu(q) \geq \eta$ (Proposition 5). For this sequence of hyper-rectangles, $k \leq \log_2(1/\eta)$.

Then by Proposition 6,

$$\left| \langle f \rangle_q - \langle f \rangle_{[0,1]^d} \right| \leq 2 \log_2(1/\eta) \|f\|. \quad (\text{A.14})$$

Thus by definition of the functions m_t, n_t for cubes with no observations, for a cube q such that $n_t(q) = 0$,

$$\begin{aligned} \left| \langle f \rangle_q - m_t(q) \right| &\stackrel{\textcircled{1}}{=} \left| \langle f \rangle_q \right| \stackrel{\textcircled{2}}{=} \left| \langle f \rangle_q - \langle f \rangle_{[0,1]^d} \right| \\ &\stackrel{\textcircled{3}}{\leq} 2 \log_2(1/\eta) \|f\| \stackrel{\textcircled{4}}{\leq} \frac{\Psi}{\sqrt{\max(1, n_t(q))}} \stackrel{\textcircled{5}}{\leq} \frac{(\Psi + D_{\mathcal{E}}) \sqrt{2 \log(2T^2/\epsilon)}}{\sqrt{\max(1, n_t(q))}}, \end{aligned}$$

where $\textcircled{1}$ is due to $m_t(q) = 0$ when $n_t(q) = 0$ by definition, $\textcircled{2}$ is from Assumption 1 ($\langle f \rangle_{[0,1]^d} = 0$), $\textcircled{3}$ is from (A.14), and $\textcircled{4}$ is from $2 \log_2(1/\eta) \leq \Psi$ (Eq. 2.49) and $n_t(q) = 0$. Recall we assume $\|f\| = 1$ for cleaner representation. We have finished the proof of Lemma 3. \square

A.2.1 Proof of Proposition 5

Proposition 5. Following (2.50), the maximal number of cubes produces by Algorithm 3 is $\frac{1}{\eta}$. The minimal cube measure is at least η .

Proof. This proposition is an immediate consequence of our partition refinement rule (2.50). The cube measures cannot be smaller than η . Otherwise, the RHS of the rule (2.50) will be nonpositive and no more splits will happen. \square

A.2.2 Proof of Proposition 6

Proposition 6 is a property of BMO functions, and can be found in textbooks (e.g., [SM93]).

Proposition 6. *For a function $f \in BMO(\mathbb{R}^d, \mu)$, and rectangles q_0, q_1, \dots, q_k such that $q_0 \subseteq q_1 \subseteq q_2 \subseteq \dots \subseteq q_k$, and a constant $K \geq 1$ such that $K\mu(q_i) \geq \mu(q_{i+1})$ for all $i \in [0, k-1]$, we have*

$$\left| \langle f \rangle_{q_0} - \langle f \rangle_{q_k} \right| \leq Kk \|f\|.$$

Proof. The proof is a consequence of basic properties of BMO function. For any two regular rectangles q_i and q_{i+1} ($i = 0, 1, 2, \dots, k-1$),

$$\begin{aligned} \left| \langle f \rangle_{q_i} - \langle f \rangle_{q_{i+1}} \right| &= \left| \frac{1}{\mu(q_i)} \int_{q_i} f d\mu - \langle f \rangle_{q_{i+1}} \right| \\ &= \left| \frac{1}{\mu(q_i)} \int_{q_i} (f - \langle f \rangle_{q_{i+1}}) d\mu \right| \\ &\leq \frac{1}{\mu(q_i)} \int_{q_i} |f - \langle f \rangle_{q_{i+1}}| d\mu \\ &\leq \frac{K}{\mu(q_{i+1})} \int_{q_i} |f - \langle f \rangle_{q_{i+1}}| d\mu && \text{(A.15)} \\ &\leq \frac{K}{\mu(q_{i+1})} \int_{q_{i+1}} |f - \langle f \rangle_{q_{i+1}}| d\mu && \text{(A.16)} \\ &\leq K \|f\|, \end{aligned}$$

where (A.15) uses $K\mu(q_i) \geq \mu(q_{i+1})$ and (A.16) uses $q_i \subseteq q_{i+1}$. Next, we use the triangle inequality and repeat the above inequality k times to get

$$\left| \langle f \rangle_{q_0} - \langle f \rangle_{q_k} \right| \leq \sum_{i=1}^k \left| \langle f \rangle_{q_{i-1}} - \langle f \rangle_{q_i} \right| \leq Kk \|f\|.$$

□

A.3 Proof of Lemma 4

Lemma 4. *For any partition \mathcal{Q} of $[0, 1]^d$, there exists $q \in \mathcal{Q}$, such that*

$$f^\delta \leq \langle f \rangle_q + \log(\mu(q)/\eta), \quad \text{(A.17)}$$

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

Proof. We use f^δ and S^δ as in Lemma 2.

Suppose, in order to get a contradiction, that for every cube $q \in \mathcal{Q}$, (A.17) is violated.

Define

$$\begin{aligned} S(q) &:= \left\{ a \in q : f(a) > \langle f \rangle_q + \log(\mu(q)/\eta) \right\}, \\ \tilde{S}(q) &:= \left\{ a \in q : f(a) > f^\delta \right\}. \end{aligned}$$

Suppose the lemma statement is false. For all $q \in \mathcal{Q}$, $f^\delta > \langle f \rangle_q + \log(\mu(q)/\eta)$. Thus we have for all $q \in \mathcal{Q}$,

$$\tilde{S}(q) \subseteq S(q).$$

We have, by the John-Nirenberg inequality,

$$\mu(S(q)) \leq \mu \left(\left\{ a \in q : |f(a) - \langle f \rangle_q| > \log(\mu(q)/\eta) \right\} \right) \leq \eta.$$

Since \mathcal{Q} is a partition (of $[0, 1]^d$), we have

$$\mu(\cup_{q \in \mathcal{Q}} S(q)) = \sum_{q \in \mathcal{Q}} \mu(S(q)) \leq \sum_{q \in \mathcal{Q}} \eta = |\mathcal{Q}|\eta.$$

On the other hand, by definition of f^δ and disjointness of the sets $\tilde{S}(q)$, we have

$$\mu(\cup_{q \in \mathcal{Q}} \tilde{S}(q)) = \mu(S^\delta) = \delta.$$

Since $\delta > |\mathcal{Q}|\eta$, we have

$$\mu(\cup_{q \in \mathcal{Q}} \tilde{S}(q)) > \mu(\cup_{q \in \mathcal{Q}} S(q)),$$

which is a contradiction to $\tilde{S}(q) \subset S(q)$ for all q . This finishes the proof. \square

A.4 Proof of Theorem 1

Theorem 1. 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 3 up to time T is

$$\sum_{t=1}^T r_t^\delta \leq \tilde{\mathcal{O}} \left(\sqrt{T|\mathcal{Q}_T|} \right), \quad (\text{A.18})$$

where \mathcal{Q}_T is the cardinality of \mathcal{Q}_T .

Proof. Under the “good event” $\mathcal{E}^{good} := \left(\bigcap_{t=1}^T \mathcal{E}(Q_t)\right) \cap \left(\bigcap_{t=1}^T \mathcal{E}(q_t^{\max})\right)$, we continue from (2.60) and get

$$\sum_{t=1}^T r_t^\delta \leq \sum_{t=1}^T 3 \frac{(\Psi + D_{\mathcal{E}}) \sqrt{2 \log(2T^2/\epsilon)}}{\sqrt{\tilde{n}_t(Q_t)}} \quad (\text{A.19})$$

$$\leq 3(\Psi + D_{\mathcal{E}}) \sqrt{2 \log(2T^2/\epsilon)} \sqrt{T} \cdot \sqrt{\sum_{t=1}^T \frac{1}{\max(1, n_{t-1}(Q_t))}} \quad (\text{A.20})$$

$$\leq 3(\Psi + D_{\mathcal{E}}) \sqrt{2 \log(2T^2/\epsilon)} \sqrt{T} \cdot \sqrt{e |\mathcal{Q}_T| \log \left(1 + (e-1) \frac{T}{|\mathcal{Q}_T|}\right)} \quad (\text{A.21})$$

where (A.19) uses (2.60), where (A.20) uses the Cauchy-Schwarz inequality, (A.21) uses (2.6).

What remains is to determine the probability under which the “good event” happens. By Lemma 3 and a union bound, we know that the event \mathcal{E}^{good} happens with probability at least $1 - 2\epsilon$. \square

A.5 Proof of Proposition 2

Proposition 2. *At any episode t , the collection of parent cubes forms a partition of the arm space.*

Proof. We first argue that any two parent cubes do not overlap. By definition, all parent cubes are dyadic cubes. By definition of dyadic cubes (2.44), two different dyadic cubes Q and Q' such that $Q \cap Q' \neq \emptyset$ must satisfy either (i) $Q' \subseteq Q$ or (ii) $Q \subseteq Q'$. From the definition of parent cubes and pre-parent cubes, we know a parent cube cannot contain another parent cube. Thus for two parent cubes Q and Q' , $Q \cap Q' \neq \emptyset$ implies $Q = Q'$. Thus two different parent cubes cannot overlap.

We then argue that the union of all parent cubes is the whole arm space. We consider the following cases for this argument. Consider any pre-parent cube Q . (1) If Q is already a parent cube, then it is obviously contained in a parent cube (itself). (2) At time episode

t , if $Q \in \mathcal{Q}_t$ is a pre-parent cube but not a parent cube, then by definition it is contained in another pre-parent cube Q_1 . If Q_1 is a parent cube, then Q is contained in a parent cube. If Q_1 is not a parent cube yet, then Q_1 is contained in another pre-parent cube Q_2 . We repeat this argument until we reach $[0, 1]^d$ which is a parent cube as long as it is a pre-parent cube. For the boundary case when $[0, 1]^d$ is a terminal cube, it is also a parent cube by convention. Therefore, any pre-parent cube is contained in a parent cube.

Next, by definition of pre-parent cubes and the zooming rule, any terminal cube is contained in a pre-parent cube. Thus any terminal cube is contained in a parent cube.

Since terminal cubes cover the arm space by definition, the parent cubes cover the whole arm space. \square

A.6 Proof of Proposition 3

Proposition 3. *Following the Zooming Rule (2.63), we have*

1. *Each parent cube of measure μ is played at most $\frac{2(\Psi+D_\varepsilon)^2 \log(2T^2/\varepsilon)}{\alpha^2 [\log(\mu/\eta)]^2}$ episodes.*
2. *Under event $\tilde{\mathcal{E}}_T$, each parent cube Q_t selected at episode t is a subset of*

$$\mathcal{X}_\delta \left((1 + 2\alpha) \log(M_d \mu(Q_t) / \eta) \right).$$

Proof. For item 1, every time a parent cube Q of measure μ is selected, all M_d of its direct sub-cubes are played. The direct sub-cubes are of measure $\frac{\mu}{M_d}$, and each such cube can be played at most $\frac{2(\Psi+D_\varepsilon)^2 \log(2T^2/\varepsilon)}{\alpha^2 \left[\log\left(\frac{\mu}{\eta}\right) \right]^2}$ times. Beyond this number, rule (2.63) will be violated, and all the direct sub-cubes can no longer be terminal cubes. Thus Q will no longer be a parent cube (since Q is no longer a pre-parent cube), and is no longer played.

Item 2 is a rephrasing of (2.69). Assume that event $\tilde{\mathcal{E}}_T = \left(\bigcap_{t=1}^T \mathcal{E}_t(q_t^{\max}) \right) \cap \left(\bigcap_{t=1}^T \mathcal{E}_t(Q_t) \right)$ is true. Let Q_t be the parent cube for episode t . By (2.58), we know, under event $\tilde{\mathcal{E}}_T$, there exists a “good” parent cube q_t^{\max} such that

$$f^\delta \leq m_t(q_t^{\max}) + H_t(q_t^{\max}) + J(q_t^{\max}).$$

By the concentration result in Lemma 3, we have, under event $\tilde{\mathcal{E}}_T$,

$$\langle f \rangle_{Q_t} \geq m_t(Q_t) - H_t(Q_t).$$

Combining the above two inequalities gives

$$\begin{aligned} f^\delta - \langle f \rangle_{Q_t} &\leq m_t(q_t^{\max}) + H_t(q_t^{\max}) + J(q_t^{\max}) - m_t(Q_t) + H_t(Q_t) \\ &\leq m_t(Q_t) + H_t(Q_t) + J(Q_t) - m_t(Q_t) + H_t(Q_t) \end{aligned} \quad (\text{A.22})$$

$$\begin{aligned} &\leq J(Q_t) + 2H_t(Q_t) \\ &\leq (1 + 2\alpha) \log(M_d \mu(Q_t)/\eta), \end{aligned} \quad (\text{A.23})$$

where (A.22) uses $U_t(Q_t) \geq U_t(q_t^{\max})$ by optimistic nature of the algorithm, and (A.23) uses rule (2.64). \square

A.7 Elaboration of Remark 6

Remark 6. Consider the (unbounded, BMO) function $f(x) = 2 \log \frac{1}{x}$, $x \in (0, 1]$. Pick $T \geq 20$. For some $t \leq T$, the t -step δ -regret of Algorithm 4 is $\mathcal{O}(\text{poly-log}(t))$ while allowing $\delta = \mathcal{O}(1/T)$ and $\eta = \Theta(1/T^4)$. Intuitively, Algorithm 4 gets close to f^δ even if f^δ is very large.

Firstly, recall the zooming number is defined as

$$\tilde{N}_{\delta, \eta, \alpha} := \sup_{\lambda \in \left(\eta^{\frac{1}{d}}, 1\right]} N_\delta \left((1 + 2\alpha) \log \left(M_d \lambda^d / \eta \right), \lambda \right). \quad (\text{A.24})$$

While this number provide a regret bound, it might overkill by allowing λ to be too small.

We define a refined zooming number

$$\tilde{N}'_{\delta, \eta, \alpha} := \sup_{\lambda \in (l_{\min}, 1]} N_\delta \left((1 + 2\alpha) \log \left(M_d \lambda^d / \eta \right), \lambda \right), \quad (\text{A.25})$$

where l_{\min} is the minimal possible cube edge length during the algorithm run. We will use this refined zooming number in this example. Before proceeding, we put forward the following claim.

Claim. Following rule (2.64), the minimal cube measure μ_{\min} at time T is at least $\Omega\left(2^{-\frac{\Psi\sqrt{2\log(2T^2/\epsilon)}}{\alpha\log 2}}\right)$.

Proof of Claim. In order to reach the minimal possible measure, we consider keep playing the cube with minimal measure (and always play a fixed cube if there are ties) and follow rule (2.64). Let t_i be the episode where i -th split happens. Since we keep playing the cube with minimal measure,

$$\frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{t_i} \approx_d \alpha \log\left(\frac{M_d 2^{-di}}{\eta}\right).$$

By taking difference between consecutive terms,

$$\frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{t_i} - \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{t_{i+1}} \approx_d \alpha \log M_d,$$

where \approx_d omits dependence on d .

Let i_{\max} be the maximal number of splits for T episodes. By using $t_0 = 1$ and $t_{i_{\max}} \leq T$, the above approximate equation gives

$$\sum_{i=0}^{i_{\max}} \left(\frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{t_i} - \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{t_{i+1}} \right) \approx_d i_{\max} \alpha \log M_d \quad (\text{A.26})$$

$$(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)} \gtrsim_d i_{\max} \alpha \log M_d, \quad (\text{A.27})$$

where the approximations omit possible dependence on d . This gives, by using $M_d = 2^d$,

$$i_{\max} \lesssim_d \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\alpha \log M_d} \leq \frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\alpha d \log 2}. \quad (\text{A.28})$$

Since each split decrease the minimal cube measure by a factor of $M_d = 2^d$, we have

$$\mu_{\min} \gtrsim 2^{-di_{\max}} \gtrsim 2^{-\frac{(\Psi + D_{\mathcal{E}})\sqrt{2\log(2T^2/\epsilon)}}{\alpha \log 2}}. \quad (\text{A.29})$$

Now we finished the proof of the claim. □

Consider the function $f(x) = 2 \log \frac{1}{x}$, $x \in (0, 1]$.

Recall

$$\mathcal{X}_\delta(\lambda) := \left\{ q \subseteq (0, 1] : \langle f \rangle_q \geq f^\delta - \lambda \right\}.$$

For this elementary decreasing function $f(x)$, we have $f^\delta = 2 \log \frac{1}{\delta}$, and $\langle f \rangle_{(0,x]} = 2 + 2 \log \frac{1}{x}$ for $x \in (0, 1]$. Thus,

$$\mathcal{X}_\delta(\lambda) = \left\{ x \in (0, 1] : \log x \leq 1 + \frac{\lambda}{2} + \log \delta \right\}.$$

By a substitution of $\lambda \leftarrow (1 + 2\alpha) \log(M_d \lambda^d / \eta)$, and using $d = 1$ and $M_d = 2$, we have

$$\mathcal{X}_\delta((1 + 2\alpha) \log(2\lambda/\eta)) = \left\{ x \in (0, 1] : x \leq e \left(\frac{2\lambda}{\eta} \right)^{\frac{1+2\alpha}{2}} \delta \right\}. \quad (\text{A.30})$$

Consider the first t ($t \leq T$) step δ -regret. For simplicity, let $t = T^\beta$ for some $\beta < 1$, $|\mathcal{Q}_t| = t$, $\eta = \frac{1}{T^4}$ and $\delta = \frac{2t}{T^4} = 2T^{\beta-4}$. We can do this since any $\delta > 0$ is f -admissible. Next we will study the zooming number $\tilde{N}_{\delta, \eta, \alpha}$ under this setting. Back to (A.30) with the above numbers,

$$\mathcal{X}_\delta((1 + 2\alpha) \log(2\lambda/\eta)) = \left\{ x \in (0, 1] : x \leq 2^{\frac{3+2\alpha}{2}} e \cdot \lambda^{\frac{1+2\alpha}{2}} T^{4\alpha+\beta-2} \right\}.$$

As an example, we take $\alpha = \frac{1}{4}$ and $\beta = \frac{1}{2}$, which gives

$$\mathcal{X}_\delta((1 + 2\alpha) \log(2\lambda/\eta)) = \left\{ x \in (0, 1] : x \leq 2^{7/4} e \lambda^{3/4} T^{-1/2} \right\}.$$

By the choice of (δ, η, α) and the claim above, for T large enough ($T \geq 20$ is sufficient), we have

$$\begin{aligned} \mu_{\min} &\gtrsim 2^{-\frac{\Psi \sqrt{2 \log(2T^2/\epsilon)}}{\alpha \log 2}} \\ &\gtrsim 2^{-(\log T)^2} \\ &\gtrsim T^{-2}, \end{aligned} \quad (\text{A.31})$$

where the last step uses $T^{-2} \leq 2^{-(\log T)^2}$ for $T \geq 20$. To bound $\tilde{N}'_{\delta, \eta, \alpha}$, we consider the following two cases.

Case I: $2^{7/4} e \cdot \lambda^{3/4} T^{-1/2} \leq 1$, i.e., $\lambda \lesssim T^{2/3}$. In this case, we need to use intervals of length λ to cover $(0, 2^{7/4} e \cdot \lambda^{3/4} T^{-1/2}]$. We need $\mathcal{O}(\lambda^{-1/4} T^{-1/2})$ intervals to cover it, which is at most $\mathcal{O}(1)$, since $\lambda \gtrsim T^{-2}$ by (A.31).

Case II: $2^{7/4} e \cdot \lambda^{3/4} T^{-1/2} > 1$, i.e., $\lambda \gtrsim T^{2/3}$. In this case, we need to use intervals of length λ to cover $(0, 1]$. We need $\mathcal{O}(\lambda^{-1})$ intervals to cover it, which is at most $\mathcal{O}(1)$, since $\lambda \gtrsim T^{2/3} \geq 1$.

In either case, we have $\tilde{N}'_{\delta,\eta,\alpha} = \mathcal{O}(1)$. Plugging back into Theorem 6 gives, with high probability, for the first $t = \sqrt{T}$ steps, the δ -regret ($\delta = \mathcal{O}(1/T)$) is of order *poly-log*(T), which is *poly-log*(t) since $T = t^2$.

A.8 Proof of Theorem 3

In this part, we provide a proof to the John-Nirenberg inequality (Theorem 3). Proofs to the John-Nirenberg inequality can be found in many textbooks on BMO functions or harmonic analysis [SM93]. Here, we present a proof by José [Mar] for completeness.

Theorem 3. (*John-Nirenberg inequality*) *Let μ be the Lebesgue measure. Let $f \in BMO(\mathbb{R}^d, \mu)$. Then there exists constants C_1 and C_2 , such that, for any hypercube $Q \subset \mathbb{R}^d$ and any $\lambda > 0$,*

$$\mu\left(\left\{x \in Q : \left|f(x) - \langle f \rangle_Q\right| \geq \lambda\right\}\right) \leq C_1 \mu(Q) \exp\left\{-\frac{\lambda}{C_2 \|f\|}\right\}.$$

Proof. The proof uses dyadic decomposition. By scaling, without loss of generality, we assume $\|f\| = 1$. Recall that μ is the Lebesgue measure. For a cube $Q \subset \mathbb{R}^d$, and $\omega > 0$, define

$$E(Q, \omega) = \left\{x \in Q : \left|f(x) - \langle f \rangle_Q\right| > \omega\right\}, \quad (\text{A.32})$$

$$\varphi(\omega) = \sup_Q \frac{\mu(E(Q, \omega))}{\mu(Q)} \quad (\text{A.33})$$

We want to show that $\varphi(\omega) \lesssim e^{-\frac{\omega}{c}}$. First take $\omega > e > 1$. Then

$$\frac{1}{\mu(Q)} \int_Q \left|f - \langle f \rangle_Q\right| d\mu \leq \|f\| = 1 \leq \omega$$

for any Q . Subdivide Q dyadically and stop when

$$\frac{1}{\mu(Q')} \int_{Q'} \left|f - \langle f \rangle_{Q'}\right| > \omega. \quad (\text{A.34})$$

Collect all such cubes (Q') to form a set $\mathcal{Q} = \{Q'_j\}_j$. Note that the cubes in \mathcal{Q} are disjoint. It could be $\mathcal{Q} = \emptyset$. Note that $\mathcal{Q} \subset \mathbb{D}_Q \setminus \{Q\}$, where \mathbb{D}_Q denotes the family of all dyadic cubes of Q .

Now we introduce the following Hardy–Littlewood type maximum M_Q , such that for a BMO function g ,

$$M_Q g(x) := \sup_{Q' \in \mathbb{D}_Q, Q' \ni x} \frac{1}{\mu(Q')} \int_{Q'} g d\mu. \quad (\text{A.35})$$

Take $g = |f - \langle f \rangle_Q|$. Then by definition of Q_j , we have

$$\{x \in Q : M_Q g(x) > \omega\} = \bigcup_{Q_j \in \mathcal{Q}} Q_j. \quad (\text{A.36})$$

For almost every $x \in E(Q, \omega)$, we have

$$\omega < |f - \langle f \rangle_Q| = g(x) \leq M_Q g(x). \quad (\text{A.37})$$

So

$$E(Q, \omega) \subset \bigcup_{Q_j \in \mathcal{Q}} Q_j \quad \text{almost everywhere.} \quad (\text{A.38})$$

Let \tilde{Q}_j be a parent cube of Q_j . Then since $\mu(\tilde{Q}_j) = 2^d \mu(Q_j)$,

$$\omega < \int_{Q_j} |f - \langle f \rangle_Q| d\mu \quad (\text{A.39})$$

$$\leq \frac{\mu(\tilde{Q}_j)}{\mu(Q_j)} \int_{\tilde{Q}_j} |f - \langle f \rangle_Q| d\mu \leq 2^d \omega. \quad (\text{A.40})$$

Thus, for $Q_j \in \mathcal{Q}$,

$$|f(x) - \langle f \rangle_Q| \leq |f(x) - \langle f \rangle_{Q_j}| + |\langle f \rangle_{Q_j} - \langle f \rangle_Q| \quad (\text{A.41})$$

$$\leq |f(x) - \langle f \rangle_{Q_j}| + \int_{Q_j} |f - \langle f \rangle_Q| d\mu \quad (\text{A.42})$$

$$\leq |f(x) - \langle f \rangle_{Q_j}| + 2^d \omega. \quad (\text{A.43})$$

Now, pick $\zeta > 2^d \omega$. For $x \in E(Q, \zeta)$, we have, for $Q_j \in \mathcal{Q}$

$$\zeta < |f(x) - \langle f \rangle_Q| \leq |f(x) - \langle f \rangle_{Q_j}| + 2^d \omega. \quad (\text{A.44})$$

Hence for $x \in E(Q, \zeta)$, $|f(x) - \langle f \rangle_{Q_j}| > \zeta - 2^d \omega$ is necessary when $|f(x) - \langle f \rangle_Q| > \omega$.

Since $\zeta > \omega$, by (A.38) we have

$$\mu(E(Q, \zeta)) = \mu(E(Q, \zeta) \cap E(Q, \omega)) \quad (\text{A.45})$$

$$\leq \sum_j \mu(E(Q, \zeta) \cap Q_j) \quad (\text{A.46})$$

$$\leq \sum_j \frac{\mu\left(\left\{x \in Q_j : \left|f(x) - \langle f \rangle_{Q_j}\right| > \zeta - 2^d \omega\right\}\right)}{\mu(Q_j)} \mu(Q_j) \quad (\text{A.47})$$

$$\leq \varphi\left(\zeta - 2^d \omega\right) \sum_j \mu(Q_j) \leq \mu(Q), \quad (\text{A.48})$$

where (A.46) is due to disjointness of Q_j and (A.38), and (A.47) uses that, for $x \in E(Q, \zeta)$,

$$\left|f(x) - \langle f \rangle_Q\right| > \omega \quad \Rightarrow \quad \left|f(x) - \langle f \rangle_{Q_j}\right| > \zeta - 2^d \omega,$$

as discussed above.

Then we have

$$\mu(E(Q, \zeta)) \leq \varphi\left(\zeta - 2^d \omega\right) \frac{1}{\omega} \sum_j \int_{Q_j} \left|f - \langle f \rangle_Q\right| d\mu \quad (\text{A.49})$$

$$\leq \frac{1}{\omega} \varphi\left(\zeta - 2^d \omega\right) \mu(Q), \quad (\text{A.50})$$

where we use (A.48) and (A.34) for (A.49), and use the definition of BMO functions and $\|f\| = 1$ for (A.50).

Hence for $\zeta > 2^d \omega$, we obtain

$$\frac{\mu(E(Q, \zeta))}{\mu(Q)} \leq \frac{1}{\omega} \varphi\left(\zeta - 2^d \omega\right). \quad (\text{A.51})$$

By taking supremum over Q on the left-hand-on of the above equation, we have

$$\varphi(\zeta) \leq \frac{\varphi\left(\zeta - 2^d \omega\right)}{\omega}. \quad (\text{A.52})$$

Put $\omega = e$. Note that $\varphi(\zeta) \leq 1$ for all $\zeta > 0$ by Definition in (A.33). Then for $0 < \zeta \leq e \cdot 2^d$, we have

$$\varphi(\zeta) \leq e \cdot e^{-\frac{\zeta}{2^d e}}. \quad (\text{A.53})$$

The above statement is true by the proof of contradiction. Assume

$$\varphi(\zeta) > e \cdot e^{-\frac{\zeta}{2^d e}} = e^{1-\frac{\zeta}{2^d e}} \quad (\text{A.54})$$

Since for all $\zeta > 0$, $\varphi(\zeta) \leq 1$, we have $1 - \frac{\zeta}{2^d e} < 0$ always true. This implies

$$\zeta > e \cdot 2^d \quad (\text{A.55})$$

This is to say if $\zeta > 0$, then $\zeta > e \cdot 2^d$. Hence (A.54) implies the domain $\zeta \in (-\infty, 0] \cup (e \cdot 2^d, +\infty)$. This shows (A.53).

Next, note that

$$(0, \infty) = (0, e \cdot 2^d] \cup \left[\bigcup_{k=1}^{\infty} (e \cdot 2^{d+k-1}, e \cdot 2^{d+k}] \right]. \quad (\text{A.56})$$

So for $e \cdot 2^d < \zeta \leq e \cdot 2^{d+1}$, $\varphi(\zeta) \leq e \cdot e^{-\frac{\zeta}{2^d e}}$. Since we have,

$$\varphi(\zeta) \leq \frac{1}{e} \varphi(\zeta - e \cdot 2^d), \quad e \cdot 2^{d+1} < \zeta \leq e \cdot 2^{d+1}. \quad (\text{A.57})$$

We see that $\varphi(\zeta - e \cdot 2^d) \leq e \cdot e^{-\frac{(\zeta - 2^d e)}{2^d e}}$ for $\zeta > e \cdot 2^d$. Hence, for $e \cdot 2^d < \zeta < e \cdot 2^{d+1}$, we have $\varphi(\zeta) \leq e \cdot e^{-\frac{\zeta}{2^d e}}$. Iterate this procedure, and we obtain the desired claim, which is, $\forall \zeta > 0$,

$$\frac{\mu(E(Q, \zeta))}{\mu(Q)} \leq e \cdot e^{-\frac{\zeta}{2^d e}} \quad \text{for every cube } Q. \quad (\text{A.58})$$

□

Appendix B

Supplementary Materials for Chapter 3

B.1 Additional Details for the Stochastic Setting

B.1.1 Concentrations of Estimators

Note. In the stochastic setting, the graph is time-invariant. For easier reference, we define random variables Z_v such that Z_v has the same distribution as $\mathcal{L}(\mathcal{P}_{t,v})$ for all t . Then, in the stochastic setting,

$$\Delta_v = \mathbb{E}[Z_{v^*}] - \mathbb{E}[Z_v], \quad \text{where } v^* \in \arg \max_v \mathbb{E}[Z_v]. \quad (\text{B.1})$$

In this part, we derive concentration bounds for hitting time estimators $\tilde{Z}_{v,n}$. To start with, we show that the hitting times are sub-exponential.

Lemma 12. *For any $v \in [K]$, $i = 1, 2, \dots$ and any integer $x > 0$, we have*

$$\mathbb{P}(Y_{v,k_{v,i}} \geq x) \leq \frac{\rho^x}{1 - \rho}.$$

where ρ is defined in Assumption 2.

Proof. Let M be the transition matrix among transient nodes. Since for any (v, i, x) ,

$$\begin{aligned} & \{Y_{v,k_{v,i}} \geq x\} \\ &= \{\text{a random walk starting from } v \text{ does not reach the absorbing node in } x \text{ steps}\}, \end{aligned}$$

we have,

$$\begin{aligned} \mathbb{P}(Y_{v,k_{v,i}} \geq x) &\leq \mathbb{P}(\{\text{random walk starting from } v \text{ does not terminate in } x \text{ steps}\}) \\ &= \sum_{h=x}^{\infty} \mathbb{P}(\{\text{random walk starting from } v \text{ terminates at step } h\}). \quad (\text{B.2}) \end{aligned}$$

Writing out the probability in (B.2) gives

$$\mathbb{P}(Y_{v,k_v,i} \geq x) \leq \sum_{l=x}^{\infty} \sum_{j=1}^K [M^l]_{ij} \leq \sum_{l=x}^{\infty} \|M^l\|_{\infty} \leq \sum_{l=x}^{\infty} \rho^l \leq \frac{\rho^x}{1-\rho}.$$

□

For the concentration results, we use Lemmas 13 and 14.

Lemma 13 (Proposition 34 by [TV15]). *Consider a martingale sequence X_1, X_2, \dots adapted to filtration $\mathcal{F}_1, \mathcal{F}_2, \dots$. For constants $c_1, c_2, \dots < \infty$, we have*

$$\mathbb{P}\left(|X_n - X_0| > \lambda \sqrt{\sum_{i=1}^n c_i^2}\right) \leq 2 \exp\left(-\frac{\lambda^2}{2}\right) + \sum_{i=1}^n \mathbb{P}(|X_i - X_{i-1}| > c_i). \quad (\text{B.3})$$

Lemma 13 is an extension of the Azuma's inequality with an extra term bounding the probability of any term in the martingale difference sequence being unbounded.

Lemma 14. *For any transient node $v \in [K]$, if $N_t^+(v) > 0$, we have*

$$\mathbb{P}\left(\left|\tilde{Z}_{v,N_t^+(v)} - \mathbb{E}[\mathcal{L}(\mathcal{P}_{s,v})]\right| \geq \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}}\right) \leq 3t^{-4}, \quad \forall s, t \in \mathbb{N}_+ \quad (\text{B.4})$$

where $\xi_t = \max\left\{1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log \rho} + \frac{5 \log t}{\log 1/\rho}\right\}$.

Proof. By Lemma 12, we have, when $x \geq \mathbb{E}[Y_{v,k_v,i}]$,

$$\begin{aligned} \mathbb{P}(|Y_{v,k_v,i} - \mathbb{E}[Y_{v,k_v,i}]| \geq x) &\leq \mathbb{P}(Y_{v,k_v,i} - \mathbb{E}[Y_{v,k_v,i}] \geq x) + \mathbb{P}(-Y_{v,k_v,i} + \mathbb{E}[Y_{v,k_v,i}] \geq x) \\ &\leq \mathbb{P}(Y_{v,k_v,i} \geq x) + \mathbb{P}(Y_{v,k_v,i} \leq \mathbb{E}[Y_{v,k_v,i}] - x) \\ &\hspace{15em} (\text{the second term is zero.}) \\ &\leq \frac{\rho^x}{1-\rho}. \end{aligned} \quad (\text{B.5})$$

Also by Lemma 12, $\mathbb{E}[Y_{v,k_v,i}] = \sum_{x=0}^{\infty} \mathbb{P}(Y_{v,k_v,i} \geq x) \leq 1 + \sum_{x=1}^{\infty} \frac{\rho^x}{1-\rho} \leq 1 + \frac{\rho}{(1-\rho)^2}$.

Since (B.5) is true only for $x \geq \mathbb{E}[Y_{v,k_v,i}]$, we have, by setting

$$\xi_t = \max\left\{1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log \rho} + \frac{5 \log t}{\log 1/\rho}\right\},$$

$$\mathbb{P}(|Y_{v,k_v,i} - \mathbb{E}[Y_{v,k_v,i}]| \geq \xi_t) \leq t^{-5}. \quad (\text{B.6})$$

Applying Lemma 13 gives

$$\begin{aligned} & \mathbb{P}\left(\left|\sum_{i=1}^{N_t^+(v)} [Y_{v,k_v,i} - \mathbb{E}[Y_{v,k_v,i}]]\right| \geq \lambda \sqrt{\sum_{i=1}^{N_t^+(v)} \xi_t}\right) \\ & \leq 2 \exp\left(\frac{-\lambda^2}{2}\right) + \sum_{i=1}^{N_t^+(v)} \mathbb{P}(|Y_{v,k_v,i} - \mathbb{E}[Y_{v,k_v,i}]| \geq \xi_t) \\ & \leq 2 \exp\left(\frac{-\lambda^2}{2}\right) + t^{-4}. \end{aligned} \quad (\text{B.7})$$

At time t , we set $\lambda = \sqrt{8 \log t}$, and from (B.7) we get

$$\begin{aligned} & \mathbb{P}\left(\left|\tilde{Z}_{v,N_t^+(v)} - \mathbb{E}[\tilde{Z}_{v,N_t^+(v)}]\right| \geq \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}}\right) \\ & = \mathbb{P}\left(\left|\sum_{i=1}^{N_t^+(v)} [Y_{v,k_v,i} - \mathbb{E}[Y_{v,k_v,i}]]\right| \geq \sqrt{8 \log t} \sqrt{\sum_{i=1}^{N_t^+(v)} \xi_t}\right) \leq 3t^{-4}. \end{aligned}$$

We now conclude the proof by using $\mathbb{E}[\tilde{Z}_{v,N_t^+(v)}] = \mathbb{E}[\mathcal{L}(\mathcal{P}_{s,v})]$ for all $s, t \in \mathbb{N}_+$. \square

B.1.2 Proof of Theorem 9

Firstly, we bound the step regret at time t by the confidence radius at time t , as is common in bandit regret analysis (e.g., [SKKS10b]).

Lemma 15. *With probability at least $1 - \frac{2}{t^4}$, Algorithm 5 satisfies, for any $t \in \mathbb{N}$,*

$$\mathbb{E}[Z_{v^*}] - \mathbb{E}[Z_{J_t}] \leq \mathcal{O}\left(\frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t^+(J_t)}}\right), \quad (\text{B.8})$$

where \mathcal{O} omits absolute constants and logarithmic factors in problem intrinsics.

Proof. By Lemma 14, with probability at least $1 - \frac{2}{t^4}$, we have

$$\mathbb{E}[Z_{J_t}] \leq \tilde{Z}_{J_t, N_t^+(J_t)} + \tilde{C}_{N_t^+(J_t), t} \quad \text{and} \quad \mathbb{E}[Z_{v^*}] \geq \tilde{Z}_{v^*, N_t^+(v^*)} - \tilde{C}_{N_t^+(v^*), t} \quad (\text{B.9})$$

Thus, with probability at least $1 - \frac{2}{t^4}$,

$$\mathbb{E}[Z_{J_t}] - \mathbb{E}[Z_{v^*}] = \tilde{Z}_{J_t, N_t^+(J_t)} + \tilde{C}_{N_t^+(J_t), t} - \left(\tilde{Z}_{v^*, N_t^+(v^*)} + \tilde{C}_{N_t^+(v^*), t} \right) \quad (\text{B.10})$$

$$\leq \tilde{Z}_{J_t, N_t^+(J_t)} + \tilde{C}_{N_t^+(J_t), t} - \left(\tilde{Z}_{J_t, N_t^+(J_t)} + \tilde{C}_{N_t^+(J_t), t} \right) \quad (\text{B.11})$$

$$\leq 2\tilde{C}_{N_t^+(J_t), t} \quad (\text{B.12})$$

$$\leq \mathcal{O} \left(\frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t^+(J_t)}} \right), \quad (\text{B.13})$$

where (B.11) uses that $J_t \in \arg \max_{v \in V} \left[\tilde{Z}_{J_t, N_t^+(v)} + \tilde{C}_{N_t^+(v), t} \right]$. \square

Theorem 9. *Let T be any positive integer. Under Assumption 2, Algorithm 5 admits regret of order*

$$\text{Reg}(T) = \tilde{\mathcal{O}} \left(\min \left\{ \frac{1}{(1-\rho)^2} \sqrt{\frac{T}{\alpha}}, \frac{1}{(1-\rho)^2} \sqrt{mT} \right\} \right),$$

where $\alpha = \min_{v \in V} \alpha_v$, and α_v is defined in Definition 5, and $\tilde{\mathcal{O}}$ omits poly-logarithmical factors in T .

In this bound the dependence on optimality gaps Δ_v are removed and all problem intrinsics are global. In other words, Algorithm 5 achieves this regret rate no matter how identical the nodes are.

Proof. Part I: $\text{Reg}(T) = \tilde{\mathcal{O}} \left(\frac{1}{(1-\rho)^2} \sqrt{\frac{T}{\alpha}} \right)$. For any $t, T \in \mathbb{N}$, consider events

$$\tilde{\mathcal{E}}_t := \left\{ \mathbb{E}[Z_{J_t}] \leq \tilde{Z}_{J_t, N_t^+(J_t)} + \tilde{C}_{N_t^+(J_t), t} \quad \text{and} \quad \mathbb{E}[Z_{v^*}] \geq \tilde{Z}_{v^*, N_t^+(v^*)} - \tilde{C}_{N_t^+(v^*), t} \right\},$$

$$\mathcal{E} := \left\{ N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -\sqrt{t \log(mT)} \quad \forall t \in [T], v \in V \right\}.$$

By Lemmas 14 and 5,

$$\mathbb{P} \left(\tilde{\mathcal{E}}_t \cap \mathcal{E} \right) \geq 1 - \frac{4}{t^4} - \frac{1}{T}. \quad (\text{B.14})$$

By Lemma 12,

$$\mathbb{E}[Z_{v^*}] = \sum_{x=0}^{\infty} \mathbb{P}(Z_{v^*} > x) \leq 1 + \frac{\rho}{(1-\rho)^2}. \quad (\text{B.15})$$

Thus, by (B.14), (B.15), we have

$$\begin{aligned}
\text{Reg}(T) &= \sum_{t=1}^T (\mathbb{E}[Z_{v^*}] - \mathbb{E}[Z_{J_t}]) \\
&\leq \left\lceil \frac{\log T}{\min_v \alpha_v^2} \right\rceil \cdot \mathbb{E}[Z_{v^*}] + \sum_{t=\left\lceil \frac{\log T}{\min_v \alpha_v^2} \right\rceil}^T \mathbb{E} \left[Z_{v^*} - Z_{J_t} \middle| \tilde{\mathcal{E}}_t \cap \mathcal{E} \right] \mathbb{P}(\tilde{\mathcal{E}}_t \cap \mathcal{E}) \\
&\quad + \mathbb{E} \left[\sum_{t=1}^T (Z_{v^*} - Z_{J_t}) \middle| \overline{\tilde{\mathcal{E}}_t \cap \mathcal{E}} \right] (1 - \mathbb{P}(\tilde{\mathcal{E}}_t \cap \mathcal{E})) \\
&\leq \left\lceil \frac{\log T}{\min_v \alpha_v^2} \right\rceil \cdot \frac{\rho}{(1-\rho)^2} + \sum_{t=\left\lceil \frac{\log T}{\min_v \alpha_v^2} \right\rceil}^T \mathcal{O} \left(\frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t^+(J_t)}} \right) \\
&\quad + \frac{\rho^2}{(1-\rho)^2} \sum_{t=1}^T \left(\frac{4}{t^4} + \frac{1}{T} \right) \quad (\text{by Lemma 15 and (B.14)}) \\
&\leq \tilde{\mathcal{O}} \left(\sum_{t=1}^T \frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t^+(J_t)}} \right), \\
&\leq \tilde{\mathcal{O}} \left(\sum_{t=1}^T \frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t(J_t) + \alpha_{J_t}(t - N_t(J_t)) - \sqrt{t \log(mT)}}} \right), \quad (\text{under event } \mathcal{E}) \\
&\leq \tilde{\mathcal{O}} \left(\sum_{t=1}^T \frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{\alpha_{J_t} t - \sqrt{t \log(mT)}}} \right) \\
&\leq \tilde{\mathcal{O}} \left(\frac{1}{(1-\rho)^2} \sqrt{\frac{T}{\alpha}} \right),
\end{aligned}$$

where $\alpha := \min_{v \in V} \alpha_v$.

Part II: $\text{Reg}(T) = \tilde{\mathcal{O}} \left(\frac{1}{(1-\rho)^2} \sqrt{mT} \right)$.

By definitions in (3.3), we know $N_t^+(v) \geq N_t(v)$ for all $t \in \mathbb{N}$ and $v \in V$. From Lemma 15, we have

$$\text{Reg}(T) \leq \sum_{t=1}^T \mathcal{O} \left(\frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t^+(J_t)}} \right) \leq \sum_{t=1}^T \mathcal{O} \left(\frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t(J_t)}} \right). \quad (\text{B.16})$$

Let $t_{i,v}$ be the i -th time the node v is played. Let B_v be the total number of time node v is played. Then we can regroup the sum in (B.16) by

$$\sum_{t=1}^T \frac{1}{N_t(J_t)} = \sum_{v \in V} \sum_{i=1}^{B_v} \frac{1}{N_{t_{i,v}}(v)}.$$

Since $t_{i,v}$ is the i -th time v is played, we have $n_{t_{i,v}}(v) = i$. Thus we have

$$\sum_{t=1}^T \frac{1}{N_t(J_t)} = \sum_{v \in V} \sum_{i=1}^{B_v} \frac{1}{N_{t_{i,v}}(v)} = \sum_{v \in V} \sum_{i=1}^{B_v} \frac{1}{i} \leq \sum_{v \in V} (1 + \log B_v) \leq m + m \log \frac{T}{m}, \quad (\text{B.17})$$

where the last inequality uses $\sum B_v = T$ and the AM-GM inequality.

We then insert the above results into (B.16) to get

$$\begin{aligned} \text{Reg}(T) &\leq \sum_{t=1}^T \mathcal{O} \left(\frac{\text{poly-log}(t)}{(1-\rho)^2 \sqrt{N_t^+(J_t)}} \right) \\ &\leq \tilde{\mathcal{O}} \left(\frac{1}{(1-\rho)^2} \sqrt{T \sum_{t=1}^T \frac{1}{N_t(J_t)}} \right) \quad (\text{use the Cauchy-Schwarz inequality}) \\ &\leq \tilde{\mathcal{O}} \left(\frac{1}{(1-\rho)^2} \sqrt{Tm \left(1 + \log \frac{T}{m}\right)} \right) \quad (\text{use (B.17)}) \\ &\leq \tilde{\mathcal{O}} \left(\frac{1}{(1-\rho)^2} \sqrt{mT} \right), \end{aligned}$$

which concludes this part. \square

B.1.3 Proof of Theorem 10

Theorem 10. *On a problem instance that satisfies Assumption 2, Algorithm 5 achieves constant regret of order $\tilde{\mathcal{O}} \left(\sum_{v: \Delta_v > 0} \left(\Delta_v + \frac{1}{(1-\rho)^2 \Delta_v} \right) \right)$, where $\tilde{\mathcal{O}}$ omits absolute constants and logarithmic dependence on problem intrinsics.*

Proof. First we define

$$T_{\min,v}^{(2)} := \min \left\{ t \in \mathbb{N} : t\alpha_v - \sqrt{t \log t} \geq \frac{32\xi_t \log t}{\Delta_v^2} \right\}. \quad (\text{B.18})$$

The proof of this theorem is developed in three steps.

Step 1: When $t \geq T_{\min,v}^{(2)}$, $\tilde{C}_{N_t^+(v),t} \leq 2\Delta_v$ with high probability.

For any $t \in \mathbb{N}$ and node $v \in V$, consider the following event.

$$\mathcal{E}_{v,t} := \left\{ N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -\sqrt{t \log t} \right\}. \quad (\text{B.19})$$

By Lemma 5, we have for any $v \in V$, with probability at least $1 - \frac{1}{t^2}$,

$$N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -\sqrt{t \log t}.$$

For a sub-optimal node v , when

$$N_t^+(v) \geq \frac{32\xi_t \log t}{\Delta_v^2}, \quad (\text{B.20})$$

we have

$$\tilde{C}_{N_t^+(v),t} = \sqrt{\frac{8\xi_t \log t}{N_t^+(v)}} \leq \frac{1}{2}\Delta_v. \quad (\text{B.21})$$

For $t \geq T_{\min,v}^{(2)}$ where $T_{\min,v}^{(2)}$ is defined in (B.18), under event $\mathcal{E}_{v,t}$, we have

$$\begin{aligned} N_t^+(v) &\geq \alpha_v t + (1 - \alpha_v)N_t(v) - \sqrt{t \log t} && (\text{under event } \mathcal{E}_{v,t}) \\ &\geq \alpha_v t - \sqrt{t \log t} && (\text{since } \alpha_v \leq 1) \\ &\geq \frac{32\xi_t \log t}{\Delta_v^2}. && (\text{B.22}) \end{aligned}$$

From above, we know (B.20) is true as long as $\mathcal{E}_{v,t}$ is true. Thus for any $t \geq T_{\min,v}^{(2)}$,

$$\mathbb{P}\left(\tilde{C}_{N_t^+(v),t} \leq \frac{1}{2}\Delta_v\right) = \mathbb{P}\left(N_t^+(v) \geq \frac{32\xi_t \log t}{\Delta_v^2}\right) \geq \mathbb{P}(\mathcal{E}_{v,t}) \geq 1 - \frac{1}{t^2}.$$

In words, $\tilde{C}_{N_t^+(v),t} \leq \frac{1}{2}\Delta_v$ as long as (1) $t \geq T_{\min,v}^{(2)}$ and (2) $\mathcal{E}_{v,t}$ is true.

Step 2: After time $T_{\min,v}^{(2)}$, a sub-optimal node is played constant number of times (in expectation).

For easier reference, for any $t \in \mathbb{N}$ and any $v \in V$, we write

$$w_{v,t} := \frac{32\xi_t \log t}{\Delta_v^2}. \quad (\text{B.23})$$

In Step 1 (Eq. B.22), we have shown that for any $t \geq$, event $\mathcal{E}_{v,t}$ implies

$$N_t^+(v) \geq w_{v,t}. \quad (\text{B.24})$$

After time $T_{\min,v}^{(2)}$, we can bound the number of times a sub-optimal arm is played (in expectation) by

$$\begin{aligned}
& \mathbb{E} \left[\sum_{t=T_{\min,v}^{(2)}}^T \mathbb{I}_{[J_t=v]} \right] \\
&= \sum_{t=T_{\min,v}^{(2)}}^T \left\{ \mathbb{E} \left[\mathbb{I}_{[J_t=v]} \middle| \mathcal{E}_{v,t} \right] \mathbb{P}(\mathcal{E}_{v,t}) + \mathbb{E} \left[\mathbb{I}_{[J_t=v]} \middle| \overline{\mathcal{E}_{v,t}} \right] (1 - \mathbb{P}(\mathcal{E}_{v,t})) \right\} \\
&\leq \left(\sum_{t=T_{\min,v}^{(2)}}^T \mathbb{P} \left(J_t = v, t \geq T_{\min,v}^{(2)} \middle| \mathcal{E}_{v,t} \right) \right) + \sum_{t=1}^{\infty} \frac{1}{t^2} \\
&\leq \sum_{t=T_{\min,v}^{(2)}}^T \mathbb{P} \left(J_t = v, t \geq T_{\min,v}^{(2)} \middle| \mathcal{E}_{v,t} \right) + \frac{\pi^2}{6} \\
&\leq \sum_{t=T_{\min,v}^{(2)}}^T \mathbb{P} \left(\tilde{Z}_{v,N_t^+(v)} + \tilde{C}_{N_t^+(v),t} \geq \tilde{Z}_{v^*,N_t^+(v^*)} + \tilde{C}_{N_t^+(v^*),t}, t \geq T_{\min,v}^{(2)} \middle| \mathcal{E}_{v,t} \right) + \frac{\pi^2}{6} \quad (\text{B.25})
\end{aligned}$$

where (B.25) uses

$$\left\{ \text{a node } v \text{ is played at } t \right\} \Rightarrow \left\{ \tilde{Z}_{v,N_t^+(v)} + \tilde{C}_{N_t^+(v),t} \geq \tilde{Z}_{v^*,N_t^+(v^*)} + \tilde{C}_{N_t^+(v^*),t} \right\}.$$

We then follow the argument by [ACBF02], and bound (B.25) by allowing $N_t^+(v)$ to take any values in $[w_{v,t}, t]$ (due to Eq. B.24), and allowing $N_t^+(v^*)$ to take any values in $[1, t]$.

This gives,

$$\begin{aligned}
& \mathbb{P} \left(\tilde{Z}_{v,N_t^+(v)} + \tilde{C}_{N_t^+(v),t} \geq \tilde{Z}_{v^*,N_t^+(v^*)} + \tilde{C}_{N_t^+(v^*),t}, t \geq T_{\min,v}^{(2)} \middle| \mathcal{E}_{v,t} \right) \\
&\leq \sum_{s=w_{v,t}}^t \sum_{s^*=1}^t \mathbb{P} \left(\tilde{Z}_{v,s} + \tilde{C}_{s,t} \geq \tilde{Z}_{v^*,s^*} + \tilde{C}_{s^*,t}, t \geq T_{\min,v}^{(2)} \middle| \mathcal{E}_{v,t} \right). \quad (\text{B.26})
\end{aligned}$$

As is used by [ACBF02], when the event

$$\left\{ \tilde{Z}_{v,s} + \tilde{C}_{s,t} \geq \tilde{Z}_{v^*,s^*} + \tilde{C}_{s^*,t} \right\}$$

is true, at least one of the following three must be true:

$$\tilde{Z}_{v,s} - \tilde{C}_{s,t} \leq \mathbb{E}[Z_v], \quad (\text{B.27})$$

$$\tilde{Z}_{v^*,s^*} + \tilde{C}_{s^*,t} \geq \mathbb{E}[Z_{v^*}], \quad (\text{B.28})$$

$$\mathbb{E}[Z_{v^*}] < \mathbb{E}[Z_v] + 2\tilde{C}_{s,t}. \quad (\text{B.29})$$

By Step 1 (Eq. B.21), we know (B.29) is false for $s \geq w_{v,t}$ and $t \geq T_{\min,v}^{(2)}$. Thus one of (B.27) and (B.28) must be true. Therefore we can continue from (B.26) to get

$$\begin{aligned} & \sum_{s=w_{v,t}}^t \sum_{s^*=1}^t \mathbb{P}\left(\tilde{Z}_{v,s} + \tilde{C}_{s,t} \geq \tilde{Z}_{v^*,s^*} + \tilde{C}_{s^*,t}, t \geq T_{\min,v}^{(2)} \mid \mathcal{E}_{v,t}\right) \\ & \leq \sum_{s=w_{v,t}}^t \sum_{s^*=1}^t \left[\mathbb{P}\left(\tilde{Z}_{v,s} + \tilde{C}_{s,t} \leq \mathbb{E}[Z_v]\right) + \mathbb{P}\left(\tilde{Z}_{v,s} + \tilde{C}_{s,t} \leq \mathbb{E}[Z_v]\right) \right] \\ & \leq \sum_{s=w_{v,t}}^t \sum_{s^*=1}^t \left(\frac{3}{t^4} + \frac{3}{t^4} \right) \end{aligned} \quad (\text{B.30})$$

$$\leq \frac{6}{t^2} \quad (\text{B.31})$$

where (B.30) uses Lemma 14.

Combining (B.25) and (B.31) gives

$$\mathbb{E} \left[\sum_{t=T_{\min,v}^{(2)}}^T \mathbb{I}_{[J_t=v]} \right] \leq \sum_{t=T_{\min,v}^{(2)}}^T \frac{6}{t^2} + \frac{\pi^2}{6} \leq \frac{7\pi^2}{6}, \quad (\text{B.32})$$

which concludes Step 2.

Step 3: Up to time $T_{\min,v}^{(2)}$, a sub-optimal node v is played $\mathcal{O}\left(\text{polylog}\left(T_{\min,v}^{(2)}, \frac{1}{\Delta_v}\right)\right)$ number of times

Recall $N_t(v)$ is the number of times we play node v up to time t .

For any integer w , we have

$$\begin{aligned}
\mathbb{E}[N_T(v)] &= \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}_{[J_t=v]}\right] \\
&\leq w + \mathbb{E}\left[\sum_{t=1}^{t=T_{\min,v}^{(2)}} \mathbb{I}_{[J_t=v, N_t(v)\geq w]} + \sum_{t=T_{\min,v}^{(2)}}^T \mathbb{I}_{[J_t=v]}\right] \\
&\leq w + \mathbb{E}\left[\sum_{t=1}^{t=T_{\min,v}^{(2)}} \mathbb{I}_{[J_t=v, N_t(v)\geq w]}\right] + \frac{7\pi^2}{6} \tag{B.33}
\end{aligned}$$

$$\leq w + \sum_{t=1}^{t=T_{\min,v}^{(2)}} \sum_{s^*=1}^t \sum_{s=w}^t \mathbb{P}\left(\tilde{Z}_{v,s} + \tilde{C}_{s,t} \geq \tilde{Z}_{v^*,s^*} + \tilde{C}_{s^*,t}\right) + \frac{7\pi^2}{6}, \tag{B.34}$$

where (B.33) uses the result of Step 2, and (B.34) uses similar argument in Step 2 (Eq. B.25 - B.26).

$$\text{We set } w := \left\lceil \frac{32\xi_{T_{\min,v}^{(2)}} \log T_{\min,v}^{(2)}}{\Delta_v^2} \right\rceil, \text{ so that at time } t \leq T_{\min,v}^{(2)}, \text{ for } s \geq w,$$

$$\tilde{C}_{s,t} = \sqrt{\frac{8\xi_t \log t}{s}} \leq \frac{1}{2}\Delta_v,$$

which mean (B.29) is false.

Also, by Lemma 14, we have

$$\mathbb{P}\left[\tilde{Z}_{v,s_v} - \tilde{C}_{s_v,t} \leq \mathbb{E}[Z_v]\right] \leq \frac{3}{t^4} \quad \mathbb{P}\left[\tilde{Z}_{v^*,s^*} + \tilde{C}_{s^*,t} \geq \mathbb{E}[Z_{v^*}]\right] \leq \frac{3}{t^4}. \tag{B.35}$$

Again we continue from (B.34) and use (B.27), (B.28) and (B.29) to get

$$\begin{aligned}
&\mathbb{E}[N_T(v)] \\
&\leq \left\lceil \frac{32\xi_{T_{\min,v}^{(2)}} \log T_{\min,v}^{(2)}}{\Delta_v^2} \right\rceil \\
&\quad + \sum_{t=1}^T \sum_{s^*=1}^t \sum_{s=v}^t \left(\mathbb{P}\left[\tilde{Z}_{v,s_v} - \tilde{C}_{s_v,t} \leq \mathbb{E}[Z_v]\right] + \mathbb{P}\left[\tilde{Z}_{v^*,s^*} + \tilde{C}_{s^*,t} \geq \mathbb{E}[Z_{v^*}]\right] \right) + \frac{7\pi^2}{6} \tag{B.36}
\end{aligned}$$

$$\leq \frac{32\xi_{T_{\min,v}^{(2)}} \log T_{\min,v}^{(2)}}{\Delta_v^2} + 1 + \frac{13\pi^2}{6}, \tag{B.37}$$

where on the last line we use (B.35).

Finally, since $\text{Reg}(T) = \sum_{v \in [K]} \Delta_v \mathbb{E}[N_T(v)]$, we have

$$\text{Reg}(T) \leq \sum_{v: \Delta_v > 0} \left[\frac{32 \xi_{T_{\min, v}^{(2)}} \log T_{\min, v}^{(2)}}{\Delta_v} + \left(1 + \frac{13\pi^2}{6}\right) \Delta_v \right],$$

where $T_{\min, v}^{(2)} = \tilde{\mathcal{O}}\left(\frac{1}{\alpha_v(1-\rho)^2 \Delta_v^2}\right)$ and $\xi_{T_{\min, v}^{(2)}} = \tilde{\mathcal{O}}\left(\frac{1}{(1-\rho)^2}\right)$. \square

B.1.4 Greedy Algorithm for the Stochastic Setting

The simplest strategy is to play the node with the largest estimated hitting time. In each epoch t , we play the node that maximizes the empirical estimates of the hitting times, $\tilde{Z}_{v, N_t^+(v)}$. This strategy is formally stated in Algorithm 7.

Algorithm 7

- 1: **Input:** A set of nodes $[K]$ (and an absorbing node $*$).
- 2: **Warm up:** Play each node once to initialize. Observe trajectories.
- 3: **for** $t = 1, 2, 3, \dots$ **do**
- 4: Select J_t to start a random walk, such that

$$J_t \in \arg \max_{v \in V} \tilde{Z}_{v, N_t^+(v)},$$

where with ties broken arbitrarily.

- 5: Observe the trajectory $\mathcal{P}_{t, v_t} := \{X_{t,0}, L_{t,1}, X_{t,1}, L_{t,2}, X_{t,2}, \dots, L_{t, H_{t, v_t}}, X_{t, H_{t, v_t}}\}$.
 - Update $N_t^+(v)$ and estimates $\tilde{Z}_{t, N_t^+(v)}$ for all $v \in [K]$.
-

Theorem 13. *Suppose Assumption 2 holds. Algorithm 7 achieves a constant regret that only depends on $\alpha_v, \alpha_{v^*}, \rho$, and Δ_v : $\text{Reg}(T) \leq \tilde{\mathcal{O}}\left(\sum_{v: \Delta_v > 0} \left(\frac{1}{\min\{\alpha_v, \alpha_{v^*}\}(1-\rho)^2 \Delta_v} + \Delta_v\right)\right)$, where v^* is the optimal node (the node with maximum hitting time), and $\tilde{\mathcal{O}}$ omits absolute constants and logarithmic dependence on problem intrinsics.*

Proof. First we define

$$T_{\min,v}^{(1)} := \min \left\{ t \in \mathbb{N} : \left(\sqrt{\frac{8\xi_t \log t}{\alpha_v t - \sqrt{t \log t}}} \leq \frac{\Delta_v}{2} \right) \wedge \left(\sqrt{\frac{8\xi_t \log t}{\alpha_{v^*} t - \sqrt{t \log t}}} \leq \frac{\Delta_v}{2} \right) \right\}, \quad (\text{B.38})$$

For a sub-optimal node v and a time $t \in \mathbb{N}$, we consider

$$\mathcal{E}_{v,t} = \left\{ N_t^+(v) - N_t(v) - (t - N_t(v))\alpha_v \geq -\sqrt{4t \log t} \right\}. \quad (\text{B.39})$$

By Lemma 5, $\mathcal{E}_{v,t}$ is true with probability at least $1 - \frac{1}{t^2}$.

For simplicity, we write $B_{n,t} = \sqrt{\frac{8\xi_t \log t}{n}}$, where $\xi_t := \max \left\{ 1 + \frac{\rho}{(1-\rho)^2}, \frac{\log(1-\rho)}{\log \rho} + \frac{5 \log t}{\log 1/\rho} \right\}$.

We have, for any sub-optimal node v , the probability of v being played at time t satisfies

$$\begin{aligned} \mathbb{P}(J_t = v) &\leq \mathbb{P} \left(\tilde{Z}_{v, N_t^+(v)} \geq \tilde{Z}_{v^*, N_t^+(v^*)} \right) && (v^* \in \arg \max_{v \in V} \mathbb{E}[Z_v]) \\ &= \mathbb{P} \left(\tilde{Z}_{v, N_t^+(v)} - \mathbb{E}[Z_v] - \left(\tilde{Z}_{v^*, N_t^+(v^*)} - \mathbb{E}[Z_{v^*}] \right) \geq \Delta_v \right) \\ &&& (\Delta_v = \mathbb{E}[Z_{v^*}] - \mathbb{E}[Z_v]) \\ &\leq \left[\mathbb{P} \left(2 \max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} > \Delta_v \right) + \mathbb{P} \left(\tilde{Z}_{v, N_t^+(v)} \geq \mathbb{E}[Z_v] + B_{N_t^+(v),t} \right) \right. \\ &&& \left. + \mathbb{P} \left(\tilde{Z}_{v^*, N_t^+(v^*)} \leq \mathbb{E}[Z_{v^*}] - B_{N_t^+(v^*),t} \right) \right], \end{aligned} \quad (\text{B.40})$$

where in (B.40) we use

$$\begin{aligned} &\left\{ \tilde{Z}_{v, N_t^+(v)} - \mathbb{E}[Z_v] - \left(\tilde{Z}_{v^*, N_t^+(v^*)} - \mathbb{E}[Z_{v^*}] \right) \geq \Delta_v \right\} \\ \implies &\left\{ 2 \max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} > \Delta_v \right\} \\ &\cup \left\{ \tilde{Z}_{v^*, N_t^+(v^*)} \geq \mathbb{E}[Z_{v^*}] + B_{N_t^+(v^*),t} \right\} \\ &\cup \left\{ \tilde{Z}_{v, N_t^+(v)} \leq \mathbb{E}[Z_v] - B_{N_t^+(v),t} \right\}, \end{aligned}$$

which can be verified by checking its contrapositive statement.

When (1) event $\mathcal{E}_{v,t}$ is true, (2) $\sqrt{\frac{8\xi_t \log t}{\alpha_v t - \sqrt{t \log t}}} \leq \frac{\Delta_v}{2}$ and (3) $\sqrt{\frac{8\xi_t \log t}{\alpha_{v^*} t - \sqrt{t \log t}}} \leq \frac{\Delta_v}{2}$, we have

$$2 \max\{B_{N_t^+(v),t}, B_{N_t^+(v^*),t}\} \leq \Delta_v.$$

Thus when $t \geq T_{\min, v}^{(1)}$, we have

$$\begin{aligned}
& \mathbb{P} \left(2 \max\{B_{N_t^+(v), t}, B_{N_t^+(v^*), t}\} \geq \Delta_v \right) \\
&= \mathbb{P} \left(2 \max\{B_{N_t^+(v), t}, B_{N_t^+(v^*), t}\} \geq \Delta_v \mid \mathcal{E}_{v, t} \right) \mathbb{P}(\mathcal{E}_{v, t}) \\
&\quad + \mathbb{P} \left(2 \max\{B_{N_t^+(v), t}, B_{N_t^+(v^*), t}\} \geq \Delta_v \mid \overline{\mathcal{E}_{v, t}} \right) [1 - \mathbb{P}(\mathcal{E}_{v, t})] \\
&\leq 1 - \mathbb{P}(\mathcal{E}_{v, t}) \leq \frac{1}{t^2}.
\end{aligned} \tag{B.41}$$

Also by Lemma 14, we have

$$\mathbb{P} \left(\tilde{Z}_{v^*, N_t^+(v^*)} \geq \mathbb{E}[Z_{v^*}] + B_{N_t^+(v^*), t} \right) \leq \frac{2}{t^2}, \quad \mathbb{P} \left(\tilde{Z}_{v, N_t^+(v)} \leq \mathbb{E}[Z_v] - B_{N_t^+(v), t} \right) \leq \frac{2}{t^2}. \tag{B.42}$$

We can now combine (B.40), (B.41) and (B.42) to get

$$\begin{aligned}
\mathbb{E}[N_T(v)] &= \sum_{t=1}^T \mathbb{P}(J_t = v) \\
&\leq T_{\min, v}^{(1)} + \sum_{t=T_{\min, v}^{(1)}}^T \mathbb{P}(v_t = v) \\
&\leq T_{\min, v}^{(1)} + \sum_{t=T_{\min, v}^{(1)}}^T \left[\mathbb{P} \left(2 \max\{B_{N_t^+(v), t}, B_{N_t^+(v^*), t}\} \geq \Delta_v \right) \right. \\
&\quad \left. + \mathbb{P} \left(\tilde{H}_{v, N_t^+(v)} \leq \mathbb{E}[H_v] - B_{N_t^+(v), t} \right) \right. \\
&\quad \left. + \mathbb{P} \left(\tilde{H}_{v^*, N_t^+(v^*)} \geq \mathbb{E}[H_{v^*}] + B_{N_t^+(v^*), t} \right) \right] \\
&\leq T_{\min, v}^{(1)} + \sum_{t=1}^{\infty} \left[\frac{1}{t^2} + \frac{2}{t^2} + \frac{2}{t^2} \right] \leq T_{\min, v}^{(1)} + \frac{5\pi^2}{6},
\end{aligned} \tag{B.43}$$

where on the last line we use (B.41) and (B.42).

Finally, we use the Wald's equation to get

$$\begin{aligned}
\text{Reg}(T) &= \sum_{t=1}^T (\mathbb{E}[\mathcal{L}(\mathcal{P}_{t, v^*})] - \mathbb{E}[\mathcal{L}(\mathcal{P}_{t, J_t})]) \\
&= \sum_{t=1}^T \sum_{v \in [K]} (\mathbb{E}[Z_{v^*}] - \mathbb{E}[Z_v]) \mathbb{P}(J_t = v) = \sum_{v \in [K], \Delta_v > 0} \Delta_v \mathbb{E}[N_T(v)].
\end{aligned}$$

From here we use $T_{\min, v}^{(1)} = \tilde{\mathcal{O}} \left(\frac{1}{\min\{\alpha_v, \alpha_{v^*}\}(1-\rho)^2 \Delta_v^2} \right)$ to conclude the proof. \square

B.2 Proofs for the Adversarial Setting

Lemma 6. For any t, i, j , it holds that $\mathbb{V}[\hat{q}_{t,ij}] = \tilde{\mathcal{O}}\left(\frac{1}{t}\right)$.

Proof. For the variance, we have

$$\mathbb{V}[\hat{q}_{t,ij}] = \sum_{m=1}^t \mathbb{V}\left[\hat{q}_{t,ij} \mid N_t^+(i) = m\right] \mathbb{P}(N_t^+(i) = m) \leq \sum_{m=1}^t \frac{1}{m} \mathbb{P}(N_t^+(i) = m) = \mathbb{E}\left[\frac{1}{N_t^+(i)}\right].$$

By Lemma 5 and a union bound, we know, for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(N_t^+(i) \geq \alpha t - \sqrt{2t \log(2TK/\delta)}, \quad \forall i \in [K], t \in [T]\right) \leq \delta.$$

Thus it holds that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{N_t^+(i)}\right] &= \mathbb{E}\left[\frac{1}{N_t^+(i)} \mid N_t^+(i) \geq \alpha t - \sqrt{2t \log(2TK/\delta)}\right] \mathbb{P}\left(N_t^+(i) \geq \alpha t - \sqrt{2t \log(2TK/\delta)}\right) \\ &\quad + \mathbb{E}\left[\frac{1}{N_t^+(i)} \mid N_t^+(i) < \alpha t - \sqrt{2t \log(2TK/\delta)}\right] \mathbb{P}\left(N_t^+(i) < \alpha t - \sqrt{2t \log(2TK/\delta)}\right) \\ &\leq \frac{1}{\max\left\{1, \alpha t - \sqrt{2t \log(2TK/\delta)}\right\}} + \delta. \end{aligned}$$

Setting $\delta = \frac{1}{T}$ concludes the proof. \square

Lemma 7. For any $\epsilon \in (0, 1)$, let

$$\mathcal{E}'_t := \left\{ |\hat{q}_{t,ij} - q_{ij}| \geq \sqrt{\frac{2\mathbb{V}[\hat{q}_{t,ij}] \log(KT/\epsilon)}{N_{t-1}^+(i)}} + \frac{\log(KT/\epsilon)}{3N_{t-1}^+(i)}, N_t^+(i) \geq \alpha t - \sqrt{t \log(2TK/\epsilon)}, \forall i, j \in [K] \right\}.$$

It holds that $\mathbb{P}(\mathcal{E}'_t) \geq 1 - \frac{2\epsilon}{T}$ and under \mathcal{E}'_t ,

$$\hat{q}_{t,ij} = q_{ij} \pm \mathcal{O}\left(\frac{\log(2TK/\epsilon)}{\alpha t}\right), \quad (\text{B.44})$$

where $\alpha = \min_{i \in [K]} \alpha_j$.

Proof. By Bennett's inequality, it holds that

$$\mathbb{P}\left(|\hat{q}_{t,ij} - q_{ij}| \geq \sqrt{\frac{2\mathbb{V}[\hat{q}_{t,ij}] \log(KT/\epsilon)}{N_{t-1}^+(i)}} + \frac{\log(KT/\epsilon)}{3N_{t-1}^+(i)}\right) \leq \epsilon \quad (\text{B.45})$$

By Lemma 5 and a union bound, we know $\mathbb{P}(\mathcal{E}'_T) \geq 1 - 2\epsilon$. By Lemma 6, we know that $\mathbb{V}[\hat{q}_{t,ij}] = \tilde{\mathcal{O}}\left(\frac{1}{t}\right)$. Thus, under event \mathcal{E}'_t , it holds that

$$\hat{q}_{t,ij} = q_{ij} \pm \tilde{\mathcal{O}}\left(\frac{\log(\epsilon^{-1})}{t}\right).$$

□

Lemma 8. For any B , let $\mathcal{E}_T(B) := \{Z_{t,j} \leq B \text{ for all } t = 1, 2, \dots, T, \text{ and } j \in [K]\}$. For any $\epsilon \in (0, 1)$ and $B = \frac{\log \frac{(1-\rho)\epsilon}{KT}}{\log \rho}$, it holds that

$$\mathbb{P}(\mathcal{E}_T(B)) \geq 1 - \epsilon,$$

and

$$\mathbb{E}[Z_{t,i} | \text{not } \mathcal{E}_T(B)] \leq KB + \frac{K}{(1-\rho)^2},$$

and

$$\mathbb{E}[Z_{t,i}^2 | \text{not } \mathcal{E}_T(B)] \leq KB^2 + \frac{2K}{(1-\rho)^3}.$$

Proof. Since all edge lengths are smaller than 1, we have, for any integer B ,

$$\begin{aligned} \mathbb{P}(Z_{t,i} > B) &\leq \mathbb{P}(\{\text{random walk starting from } i \text{ does not terminate in } B \text{ steps}\}) \\ &= \sum_{l=B}^{\infty} \mathbb{P}(\{\text{random walk starting from } i \text{ terminates at step } l\}) \\ &= \sum_{l=B}^{\infty} \sum_{j=1}^K [M^l]_{ij} \leq \sum_{l=B}^{\infty} \|M^l\|_{\infty} \leq \sum_{l=B}^{\infty} \rho^l \leq \frac{\rho^B}{1-\rho}. \end{aligned}$$

Thus with probability at least $1 - \frac{\rho^B}{1-\rho}$, we have $Z_{t,i} \leq B$. We define

$$\mathcal{E}_T(B) := \{Z_{t,j} \leq B \text{ for all } t = 1, 2, \dots, T, \text{ and } j \in [K]\}.$$

By a union bound, $\mathbb{P}(\mathcal{E}_T(B)) \geq 1 - \frac{KT\rho^B}{1-\rho}$. Now we can set $B = \frac{\log \frac{(1-\rho)\epsilon}{KT}}{\log \rho}$ so that $\mathbb{P}(\mathcal{E}_T(B)) \geq 1 - \epsilon$.

The random variables $Z_{t,i}$ also has the memorylessness-type property:

$$\begin{aligned}
& \mathbb{E}[Z_{t,i} | \text{not } \mathcal{E}_T(B)] \\
& \leq \mathbb{E}[Z_{t,i} | Z_{t,i} > B] \leq \sum_{l=B+1}^{\infty} l \frac{\mathbb{P}(\{\mathcal{P}_{t,i} \text{ terminates at step } l\} \cap \{Z_{t,i} > B\})}{\mathbb{P}(Z_{t,i} > B)} \\
& = \sum_{l=B+1}^{\infty} l \frac{\sum_j \mathbb{P}(\{\mathcal{P}_{t,i} \text{ terminates at step } l\} \cap \{\text{the } (B+1)\text{-th step is at } j\})}{\sum_j \mathbb{P}(\{\text{the } (B+1)\text{-th step is at } j\})} \\
& \leq \sum_{l=B+1}^{\infty} l \sum_j \mathbb{P}(\{\mathcal{P}_{t,i} \text{ terminates at step } l\} | \{\text{the } (B+1)\text{-th step is at } j\}) \\
& = \sum_j \sum_{l=1}^{\infty} (l+B) \mathbb{P}(\{\mathcal{P}_{t,j} \text{ terminates at step } l\}) \\
& = \sum_j \mathbb{E}[Z_{t,j} + B] \leq KB + \sum_j \mathbb{E}[Z_{t,j}]
\end{aligned}$$

where we use Markov property on the second last line.

Since $\mathbb{E}[Z_{t,j}] \leq \mathcal{O}\left(\frac{1}{(1-\rho)^2}\right)$, we insert this into the above equation to get

$$\mathbb{E}[Z_{t,i} | \text{not } \mathcal{E}_T(B)] \leq \mathcal{O}\left(KB + \frac{K}{(1-\rho)^2}\right).$$

Similarly, we have

$$\mathbb{E}[Z_{t,i}^2 | \text{not } \mathcal{E}_T(B)] \leq \mathcal{O}\left(KB^2 + \frac{K}{(1-\rho)^3}\right).$$

□

Lemma 10. *With probability at least $1 - 6\epsilon$, we have*

$$\begin{aligned}
& \sum_{t=1}^T \sum_j \frac{p_{tj}}{\hat{p}_{tj}^2} \mathbb{1}_{[j \in \mathcal{P}_{t,J_t}]} \leq \kappa T + \tilde{\mathcal{O}}\left(\sqrt{T \log(1/\epsilon)}\right), \\
& \sum_t \sum_j p_{tj} \hat{Z}_{t,j} - \sum_t \frac{l_{t,J_t} - B}{B} \leq \kappa \beta T + \mathcal{O}\left(\sqrt{(1 + \beta \kappa) T \log(1/\epsilon)}\right), .
\end{aligned}$$

Proof. To prove the first inequality, we verify that $\frac{p_{tj}}{\hat{p}_{tj}}$ is bounded, compute the (conditional) expectation of $\frac{p_{tj}}{\hat{p}_{tj}}$, and apply the Azuma's inequality. By Lemma 7, under event \mathcal{E}'_t , we

have

$$\begin{aligned}
\frac{1}{\tilde{p}_{tj}^2} &= \frac{1}{\left(p_{tj} + \sum_{i \neq j} p_{ti} \left(q_{ij} \pm \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right)\right)\right)^2} \\
&= \frac{1}{\left(\tilde{p}_{tj} \pm \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right)\right)^2} \\
&= \frac{1}{\tilde{p}_{tj}^2} + \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right), \tag{B.46}
\end{aligned}$$

where the first equation uses Lemma 7, and the last inequality uses the Taylor expansion that $\frac{1}{(x-a)^2} = \frac{1}{a^2} + \mathcal{O}\left(\frac{x}{a^3}\right)$ (with $x = \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right)$).

By (B.46), the conditional expectation of $\frac{p_{tj}}{\tilde{p}_{tj}^2}$ is

$$\begin{aligned}
\mathbb{E}\left[\frac{p_{tj}}{\tilde{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} \middle| \mathcal{E}'_t\right] &= \mathbb{E}\left[\frac{p_{tj}}{\tilde{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} \middle| \mathcal{E}'_t\right] + \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right) \quad (\text{by Eq. B.46}) \\
&\leq \frac{1}{1 - 2\epsilon/T} \mathbb{E}\left[\frac{p_{tj}}{\tilde{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} \right] + \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right) \\
&\leq \mathbb{E}[p_{tj}] + \frac{1 - \sqrt{\alpha_j}}{1 + \sqrt{\alpha_j}} + \mathcal{O}\left(\frac{\log(TK/\epsilon)}{\alpha t}\right), \tag{B.47}
\end{aligned}$$

where ① uses $\mathbb{P}(\mathcal{E}'_t) \geq 1 - \frac{2\epsilon}{T}$, and the last line uses $\frac{x}{x+a(1-x)} \leq x + \frac{1-\sqrt{a}}{1+\sqrt{a}}$ for $x \in [0, 1]$ and $a \in [0, 1]$ (Proposition 7 in Appendix B.2.2), and simply removes the slight dependence on $\frac{\epsilon}{T}$ in the first term (since there is a $\tilde{\mathcal{O}}$ residual term).

We take summation over j on both sides of (B.47), to get

$$\mathbb{E}\left[\sum_j \frac{p_{tj}}{\tilde{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} \middle| \mathcal{E}'_t\right] \leq \kappa + \mathcal{O}\left(\frac{2 \log(TK/\epsilon)}{\alpha t}\right)$$

We now apply Lemma 13 (in Appendix B.2) to $\left\{\sum_j \frac{p_{tj}}{\tilde{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} \mathbb{I}_{[\mathcal{E}'_t]}\right\}_t$ and get, with probability at least $1 - 3\epsilon$ ($\epsilon \leq T$),

$$\sum_{t=1}^T \sum_j \frac{p_{tj}}{\tilde{p}_{tj}^2} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} \leq \kappa T + \tilde{\mathcal{O}}\left(\sqrt{T \log(1/\epsilon)}\right).$$

For the second inequality in the lemma statement, we verify that $\sum_j p_{tj} \hat{Z}_{t,j} - \sum_j \frac{l_{t,J_t} - B}{B}$ is bounded, compute its (conditional) expectation, and apply the Azuma's inequality.

Firstly, it holds that $\sum_j p_{tj} \widehat{Z}_{t,j}$ is bounded conditioning on $\mathcal{E}_T(B)$:

$$\begin{aligned} \sum_j p_{tj} \widehat{Z}_{t,j} &= \sum_j p_{tj} \frac{\frac{Z_{t,j}-B}{B} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} + \beta}{\widehat{p}_{tj}} \\ &\stackrel{\textcircled{1}}{=} \sum_j p_{tj} \frac{\frac{Z_{t,j}-B}{B} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} + \beta}{\widetilde{p}_{tj}} + \sum_j \frac{\beta p_{tj}}{\widetilde{p}_{tj}} + \widetilde{\mathcal{O}}\left(\frac{\log(1/\epsilon)}{t}\right) \\ &\leq 1 + \beta\kappa + \widetilde{\mathcal{O}}\left(\frac{\log(1/\epsilon)}{t}\right), \end{aligned}$$

where $\textcircled{1}$ uses a Taylor expansion, and the last line uses Proposition 7.

Also, we have

$$\begin{aligned} \mathbb{E} \left[\sum_j p_{tj} \widehat{Z}_{t,j} \middle| \mathcal{E}'_t \right] &\stackrel{\textcircled{1}}{=} \mathbb{E} \left[\sum_j p_{tj} \frac{\frac{Z_{t,j}-B}{B} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} + \beta}{\widetilde{p}_{tj}} \middle| \mathcal{E}'_t \right] + \widetilde{\mathcal{O}}\left(\frac{\log(1/\epsilon)}{t}\right) \\ &\stackrel{\textcircled{2}}{\leq} \frac{1}{1-2\epsilon/T} \mathbb{E} \left[\sum_j p_{tj} \frac{\frac{Z_{t,j}-B}{B} \mathbb{I}_{[j \in \mathcal{P}_{t,J_t}]} + \beta}{\widetilde{p}_{tj}} \right] + \widetilde{\mathcal{O}}\left(\frac{\log(1/\epsilon)}{t}\right) \\ &\stackrel{\textcircled{3}}{\leq} \mathbb{E} \left[\sum_j p_{tj} \frac{Z_{t,j}-B}{B} + \beta \sum_j \frac{p_{tj}}{\widetilde{p}_{tj}} \right] + \widetilde{\mathcal{O}}\left(\frac{\log(1/\epsilon)}{t}\right) \\ &\stackrel{\textcircled{4}}{\leq} \mathbb{E} \left[\sum_j p_{tj} \frac{Z_{t,j}-B}{B} \right] + \beta\kappa + \widetilde{\mathcal{O}}\left(\frac{\log(1/\epsilon)}{t}\right), \end{aligned} \tag{B.48}$$

where $\textcircled{1}$ uses Lemma 7 and a Taylor expansion, $\textcircled{2}$ uses $\mathbb{P}(\mathcal{E}'_t) \geq 1 - 2\epsilon/T$, $\textcircled{3}$ simply removes the slight dependence on $\frac{\epsilon}{T}$ in the first term (since there is a $\widetilde{\mathcal{O}}$ residual term), and $\textcircled{4}$ uses Proposition 7.

Also, we have

$$\mathbb{E} [l_{t,J_t} | \mathcal{E}'_t] \geq \frac{\mathbb{E} [l_{t,J_t}] - \mathbb{E} [l_{t,J_t} | \text{not } \mathcal{E}'_t]}{\mathbb{P}(\mathcal{E}'_t)} \geq \frac{\mathbb{E} [l_{t,J_t}] - \frac{\epsilon}{T(1-\rho)}}{1 - \frac{\epsilon}{T(1-\rho)}} \geq \mathbb{E} [l_{t,J_t}] - \frac{\frac{\epsilon}{T(1-\rho)}}{1 - \frac{\epsilon}{T(1-\rho)}}.$$

From above, we know $\left\{ \left(\sum_j p_{tj} \widehat{Z}_{t,j} - \frac{l_{t,J_t}-B}{B} + \beta\kappa + \frac{\frac{\epsilon}{T(1-\rho)}}{B(1-\frac{\epsilon}{T(1-\rho)})} + \widetilde{\mathcal{O}}\left(\frac{\log(1/\epsilon)}{t}\right) \right) \mathbb{I}_{[\mathcal{E}'_t]} \right\}_t$ is a super-martingale difference sequence. We can now apply Lemma 13 (extended Azuma's inequality, in Appendix B.1.1) and get

$$\sum_t \sum_j p_{tj} \widehat{Z}_{t,j} - \sum_t \frac{l_{t,J_t}-B}{B} \leq \kappa\beta T + \mathcal{O}\left(\sqrt{(1+\beta\kappa)T \log(1/\epsilon)}\right). \tag{B.49}$$

□

B.2.1 Proof of Theorem 12

Theorem 12. Fix any $T > \sqrt{128 \log 8}$ and $\sigma < \frac{1}{7}$. On a graph of K nodes and any pair of nodes are connected with probability p , there exists $j \in [K]$ and a sequence of edge lengths, such that regret incurred by any policy satisfies

$$\mathbb{P}_{\mathfrak{J}, \pi} \left(\text{Reg}_j^{\text{adv}}(T) \geq \min \left\{ \sqrt{\frac{(1-Kp)^2 \sigma^2 T}{32p}}, \sqrt{\frac{(K-1)(1-Kp)\sigma^2 T}{32 \left(1 + \frac{p}{1-Kp}\right)}} \right\} \right) \geq \frac{1}{8}. \quad (\text{B.50})$$

Proof. A deterministic problem instance \mathfrak{J} is represented by T graphs $\mathfrak{J} = (G_1, G_2, \dots, G_T)$. For this part, the graph G_t consists of edge lengths: $G_t := \left(\{l_{i*}^{(t)}\}_{i \in [K]}, \{l_{ij}^{(t)}\}_{i, j \in [K]} \right)$, where $l_{i*}^{(t)}$ is the length from i to $*$ in G_t , and $l_{ij}^{(t)}$ is the length from i to j in G_t . A stochastic problem instance is represented by a distribution over deterministic problem instances.

By Proposition 9 (in Appendix B.2.2), it suffices to consider stochastic instances. Next we construct stochastic problem instances to prove Theorem 12.

We first sample T *i.i.d.* Gaussian random variables $\eta_t \sim \mathcal{N}(0, \sigma^2)$ (σ to be specified later). Consider the stochastic problem instance \mathfrak{J} : $\mathfrak{J} = (G_1, G_2, \dots, G_T)$. In G_t , $l_{1*}^{(t)} = \text{clip} \left(\frac{1}{2} + \frac{\epsilon}{1-Kp} + \eta_t \right)$, $l_{i*}^{(t)} = \text{clip} \left(\frac{1}{2} + \eta_t \right)$ ($i = 2, 3, \dots, K$), $l_{ij}^{(t)} = \text{clip} \left(\frac{1}{2} + \eta_t \right)$ ($i, j \in [K]$), where “clip” takes a number and clip its value to $[0, 1]$.

By this construction, the hitting times have the following properties. If no clipping happens, the hitting times $\mathbf{Z}_t = [Z_{t,1}, Z_{t,2}, \dots, Z_{t,K}]$ at time t satisfies

$$\mathbb{E}[\mathbf{Z}_t] = (1-Kp) \left(\left(\frac{1}{2} + \eta_t \right) \mathbf{1} + \frac{\epsilon}{1-Kp} \mathbf{e}_1 \right) + M \mathbb{E}[\mathbf{Z}_t] + Kp \left(\frac{1}{2} + \eta_t \right) \mathbf{1}, \quad (\text{B.51})$$

where $\mathbf{e}_1 = [1, 0, 0, \dots, 0]^\top$ and $\mathbf{1} = [1, 1, \dots, 1]^\top$. The first term on the right-hand-side of (B.51) accounts for the edge lengths (hitting time) from hitting absorbing node. The last two terms in the right-hand-side of (B.51) accounts for the edge lengths (hitting time) from remaining in the transient nodes.

This gives,

$$(I - M) \mathbb{E}[\mathbf{Z}_t] = \left(\frac{1}{2} + \eta_t \right) \mathbf{1} + \epsilon \mathbf{e}_1, \quad \text{and} \quad \mathbb{E}[\mathbf{Z}_t] = \left(I + \frac{M}{1-Kp} \right) \left(\left(\frac{1}{2} + \eta_t \right) \mathbf{1} + \epsilon \mathbf{e}_1 \right).$$

If $\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{2\epsilon}{1-Kp}\right]$, no clipping happens and $\mathbb{E}[Z_{t,1}] \geq \mathbb{E}[Z_{t,j}] + \epsilon$ for all $j = 2, 3, \dots, K$. If $\eta_t \notin \left[-\frac{1}{2}, \frac{1}{2} - \frac{2\epsilon}{1-Kp}\right]$, some edges are clipped and $\mathbb{E}[Z_{t,1}] \geq \mathbb{E}[Z_{t,j}]$. Thus we have, for all $j \geq 2$,

$$\mathbb{E}[Z_{t,1}] \geq \mathbb{E}[Z_{t,j}] + \epsilon \mathbb{I}_{\left[\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}\right]\right]}. \quad (\text{B.52})$$

Construct another instance \mathfrak{J}' . We let

$$\mathfrak{J}' = (G'_1, G'_2, \dots, G'_T)$$

and

$$G'_t = \left(\{l_{i*}^{(t)'}\}_{i \in [K]}, \{l_{ij}^{(t)'}\}_{i,j \in [K]} \right),$$

where $l_{i*}^{(t)'}$ is the length from i to $*$ in G'_t , and $l_{ij}^{(t)'}$ is the length from i to j in G'_t . We again sample $\eta'_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ (σ to be specified later). Let instance \mathfrak{J}' satisfy $l_{1*}^{(t)'} = \text{clip}\left(\frac{1}{2} + \frac{\epsilon}{1-Kp} + \eta'_t\right)$, $l_{2*}^{(t)'} = \text{clip}\left(\frac{1}{2} + \frac{2\epsilon}{1-Kp} + \eta'_t\right)$, $l_{i*}^{(t)'} = \text{clip}\left(\frac{1}{2} + \eta'_t\right)$ ($i > 2$), and $l_{ij}^{(t)'} = \text{clip}\left(\frac{1}{2} + \eta'_t\right)$ for all $i, j \in [K]$.

Let $\mathbf{Z}'_t = [Z'_{t,1}, Z'_{t,2}, \dots, Z'_{t,K}]$ be the hitting times at time t in instance \mathfrak{J}' . Thus for $j \neq 2$, in instance \mathfrak{J}' ,

$$\mathbb{E}[Z'_{t,2}] \geq \mathbb{E}[Z'_{t,j}] + \epsilon \mathbb{I}_{\left[\eta'_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{2\epsilon}{1-Kp}\right]\right]} \quad (\text{B.53})$$

From (B.52), we know, in instance \mathfrak{J} , when there are at least $\frac{3}{4}T$ unclipped rounds and node 1 is played no more than $\frac{1}{2}T$ times, then in at least $\frac{1}{4}T$ rounds, a regret of ϵ is incurred. Similarly, in instance \mathfrak{J}' , when there are at least $\frac{3}{4}T$ unclipped rounds and node 1 is played more than $\frac{1}{2}T$ times, then in at least $\frac{1}{4}T$ rounds, a regret of (at least) ϵ is incurred.

Now we define some notations to write the above observations symbolically. Let $\mathbb{P}_{\mathfrak{J},\pi}$ (resp. $\mathbb{P}_{\mathfrak{J}',\pi}$) be the probability measure on running π on \mathfrak{J} (resp. \mathfrak{J}'). Let $\mathbb{E}_{\mathfrak{J},\pi}$ (resp. $\mathbb{E}_{\mathfrak{J}',\pi}$) be the expectation with respect to $\mathbb{P}_{\mathfrak{J},\pi}$ (resp. $\mathbb{P}_{\mathfrak{J}',\pi}$). Let N_i (resp. N'_i) be the number of times i is played in \mathfrak{J} (resp. \mathfrak{J}').

Since $\mathbb{E}_{\mathfrak{J},\pi} \left[\sum_j N_j \right] = T$, there exists $i \in \{2, 3, \dots, K\}$ such that $\mathbb{E}_{\mathfrak{J},\pi} [N_i] \leq \frac{T}{K-1}$. Without loss of generality, we assume $i = 2$. Otherwise, we can rename the nodes $\{2, 3, \dots, K\}$ such that i becomes 2.

Let $\text{Reg}_1^{\text{adv}}(T)$ (resp. $\text{Reg}_2^{\text{adv}'}(T)$) be the regret in \mathfrak{J} against node 1 (resp. in \mathfrak{J}' against node 2). Let $u = \frac{1}{4}\epsilon T$. Let $W = \sum_t \mathbb{I}[\eta_t \in [-\frac{1}{2}, \frac{1}{2} + \frac{\epsilon}{1-Kp}]]$, and $W' = \sum_t \mathbb{I}[\eta'_t \in [-\frac{1}{2}, \frac{1}{2} + \frac{\epsilon}{1-Kp}]]$. Using the above notations, we have

$$\begin{aligned} & \mathbb{P}_{\mathfrak{J},\pi} \left(\text{Reg}_1^{\text{adv}}(T) \geq u \right) \\ & \geq \mathbb{P}_{\mathfrak{J},\pi} (N_1 < T/2 \quad \text{and} \quad W \geq 3T/4) \geq \mathbb{P}_{\mathfrak{J},\pi} (N_1 < T/2) - \mathbb{P}_{\mathfrak{J},\pi} (W < 3T/4) \end{aligned}$$

and

$$\begin{aligned} & \mathbb{P}_{\mathfrak{J}',\pi} \left(\text{Reg}_2^{\text{adv}'}(T) \geq u \right) \\ & \geq \mathbb{P}_{\mathfrak{J}',\pi} (N'_1 \geq T/2 \quad \text{and} \quad W' \geq 3T/4) \geq \mathbb{P}_{\mathfrak{J}',\pi} (N'_1 \geq T/2) - \mathbb{P}_{\mathfrak{J}',\pi} (W' < 3T/4), \end{aligned}$$

which gives

$$\begin{aligned} & \mathbb{P}_{\mathfrak{J},\pi} \left(\text{Reg}_1^{\text{adv}}(T) \geq u \right) + \mathbb{P}_{\mathfrak{J}',\pi} \left(\text{Reg}_2^{\text{adv}'}(T) \geq u \right) \\ & \geq \mathbb{P}_{\mathfrak{J},\pi} (N_1 < T/2) - \mathbb{P}_{\mathfrak{J},\pi} (W < 3T/4) + \mathbb{P}_{\mathfrak{J}',\pi} (N'_1 \geq T/2) - \mathbb{P}_{\mathfrak{J}',\pi} (W' < 3T/4). \quad (\text{B.54}) \end{aligned}$$

The quantities $\mathbb{P}_{\mathfrak{J},\pi} (W < 3T/4)$ and $\mathbb{P}_{\mathfrak{J}',\pi} (W' < 3T/4)$ can be easily handled since η_t are Gaussian (Proposition 8). Now we turn to lower bound $\mathbb{P}_{\mathfrak{J},\pi} (N_1 < T/2) + \mathbb{P}_{\mathfrak{J}',\pi} (N'_1 \geq T/2)$, and then select a proper ϵ to maximize this lower bound.

By the definition of total variation and the Pinsker's inequality,

$$\begin{aligned} \mathbb{P}_{\mathfrak{J},\pi} (N_1 \geq T/2) + \mathbb{P}_{\mathfrak{J}',\pi} (N'_1 < T/2) &= 1 + \mathbb{P}_{\mathfrak{J},\pi} (N_1 \geq T/2) - \mathbb{P}_{\mathfrak{J}',\pi} (N'_1 \geq T/2) \\ &\geq 1 - d_{TV} (\mathbb{P}_{\mathfrak{J},\pi}, \mathbb{P}_{\mathfrak{J}',\pi}) \\ &\geq 1 - \sqrt{2D_{KL} (\mathbb{P}_{\mathfrak{J},\pi}, \mathbb{P}_{\mathfrak{J}',\pi})}. \end{aligned}$$

Let $\mathcal{Q}_{t,j}$ (resp. $\mathcal{Q}'_{t,j}$) be the probability space generated by playing j at t in instance \mathfrak{J} (resp. \mathfrak{J}'). By chain rule, we have

$$D_{KL} (\mathbb{P}_{\mathfrak{J},\pi} \| \mathbb{P}_{\mathfrak{J},\pi}) = \sum_{t=1}^T \sum_{j \in [K]} \mathbb{P}_{\mathfrak{J},\pi} (J_t = j) D_{KL} (\mathcal{Q}_{t,j} \| \mathcal{Q}'_{t,j}). \quad (\text{B.55})$$

Let $X_0, L_1, X_1, L_2, \dots$ be the nodes and edge length of each step in the trajectory after playing a node. The sample space of $\mathcal{Q}_{t,j}$ and $\mathcal{Q}'_{t,j}$ is spanned by $X_0, L_1, X_1, L_2, \dots$.

By Markov property, we have, for all $i, j \in [K]$ and $k \in \mathbb{N}_+$,

$$\mathcal{Q}_{t,i}(L_{k+1}, X_{k+1}, L_{k+2}, X_{k+2}, \dots | X_k = j) = \mathcal{Q}_{t,j}. \quad (\text{B.56})$$

$$\mathcal{Q}'_{t,j}(L_{k+1}, X_{k+1}, L_{k+2}, X_{k+2}, \dots | X_k = j) = \mathcal{Q}'_{t,j}.$$

Thus by chain rule,

$$\begin{aligned} & D_{KL}(\mathcal{Q}_{t,i} \| \mathcal{Q}'_{t,i}) \\ &= D_{KL}(\mathcal{Q}_{t,i}(X_1, L_1) \| \mathcal{Q}'_{t,i}(X_1, L_1)) \\ & \quad + \sum_{x \in [K]} \int_0^1 \mathbb{P}(X_1 = x, L_1 = l) D_{KL}(\mathcal{Q}_{t,i}(X_2, L_2, \dots | X_1 = x, L_1 = l) \| \mathcal{Q}'_{t,i}(X_2, L_2, \dots | X_1 = x, L_1 = l)) dl \\ &= D_{KL}(\mathcal{Q}_{t,i}(X_1, L_1) \| \mathcal{Q}'_{t,i}(X_1, L_1)) + \sum_{j \in [K]} m_{ji} D_{KL}(\mathcal{Q}_{t,j} \| \mathcal{Q}'_{t,j}). \end{aligned} \quad (\text{B.57})$$

Let f be the *p.d.f.* of $\mathcal{N}(\frac{1}{2}, \sigma^2)$ truncated to $[0, 1]$. Let f^* be the *p.d.f.* of $\mathcal{N}(\frac{1}{2} + \frac{2\epsilon}{1-Kp}, \sigma^2)$ clipped to $[0, 1]$. Let ϕ (resp. Φ) be the *p.d.f.* (resp. *c.d.f.*) of the standard normal distribution. Thus we have

$$\begin{aligned} & D_{KL}(\mathcal{Q}_{t,2}(X_1, L_1) \| \mathcal{Q}'_{t,2}(X_1, L_1)) \\ &= \int_0^1 (1-Kp)f(z) \log \frac{(1-Kp)f(z)}{(1-Kp)f^*(z)} dz + K \int_0^1 pf(z) \log \frac{pf(z)}{pf(z)} dz \\ &= (1-Kp) \int_0^1 f(z) \log \frac{f(z)}{f^*(z)} dz \\ &= (1-Kp) D_{KL}\left(\mathcal{N}\left(\frac{1}{2}, \sigma^2\right) \Big|_{\text{clip}}, \mathcal{N}\left(\frac{1}{2} + \frac{2\epsilon}{1-Kp}, \sigma^2\right) \Big|_{\text{clip}}\right) \\ &\leq (1-Kp) D_{KL}\left(\mathcal{N}\left(\frac{1}{2}, \sigma^2\right), \mathcal{N}\left(\frac{1}{2} + \frac{2\epsilon}{1-Kp}, \sigma^2\right)\right) \\ &= \frac{2\epsilon^2}{(1-Kp)\sigma^2}, \end{aligned} \quad (\text{B.58})$$

where (B.58) uses monotonicity of f -divergence (e.g., [CS04]).

Also,

$$D_{KL}(\mathcal{Q}_{t,i}(X_1, L_1) \| \mathcal{Q}'_{t,i}(X_1, L_1)) = 0, \quad \text{for } i \neq 2.$$

Next, define

$$D = [D_{KL}(\mathcal{Q}_{t,1} \| \mathcal{Q}'_{t,1}), D_{KL}(\mathcal{Q}_{t,2} \| \mathcal{Q}'_{t,2}), \dots, D_{KL}(\mathcal{Q}_{t,K} \| \mathcal{Q}'_{t,K})]^\top,$$

and

$$c = [D_{KL}(\mathcal{Q}_{t,0}(X_1, L_1) \| \mathcal{Q}'_{t,0}(X_1, L_1)), \dots, D_{KL}(\mathcal{Q}_{t,K}(X_1, L_1) \| \mathcal{Q}'_{t,K}(X_1, L_1))]^\top.$$

Then we can rewrite (B.57) as

$$D = MD + c, \quad \text{and thus} \quad D = \left(I + \frac{M}{1 - Kp} \right) c.$$

Solving the above gives, for all $i \in [K]$,

$$\begin{aligned} & D_{KL}(\mathcal{Q}_{t,i} \| \mathcal{Q}'_{t,i}) \\ &= D_{KL}(\mathcal{Q}_{t,i}(X_1, L_1) \| \mathcal{Q}'_{t,i}(X_1, L_1)) + \frac{p}{1 - Kp} \sum_{j \in [K]} D_{KL}(\mathcal{Q}_{t,j}(X_1, L_1) \| \mathcal{Q}'_{t,j}(X_1, L_1)) \\ &\leq \begin{cases} \left(1 + \frac{p}{1 - Kp}\right) \frac{2\epsilon^2}{(1 - Kp)\sigma^2}, & \text{if } i = 2, \\ \frac{2p\epsilon^2}{(1 - Kp)^2\sigma^2}, & \text{otherwise.} \end{cases} \end{aligned}$$

Plugging above computation, and that $\mathbb{E}_{\mathfrak{J},\pi}[N_2] \leq \frac{T}{K-1}$, back to (B.55), we have

$$\begin{aligned} D_{KL}(\mathbb{P}_{\mathfrak{J},\pi} \| \mathbb{P}_{\mathfrak{J}',\pi}) &= \sum_{t=1}^T \sum_{j \in [K]} \mathbb{P}_{\mathfrak{J},\pi}(J_t = j) D_{KL}(\mathcal{Q}_{t,i} \| \mathcal{Q}'_{t,i}) \\ &= \sum_j \mathbb{E}_{\mathfrak{J},\pi}[N_j] D_{KL}(\mathcal{Q}_{t,i} \| \mathcal{Q}'_{t,i}) \\ &\leq \left(1 + \frac{p}{1 - Kp}\right) \frac{2\epsilon^2}{(1 - Kp)\sigma^2} \cdot \frac{T}{K-1} + \frac{2p\epsilon^2 T}{(1 - Kp)^2\sigma^2}. \end{aligned}$$

Thus by Pinsker's inequality,

$$\begin{aligned} d_{TV}(\mathbb{P}_{\mathfrak{J},\pi}, \mathbb{P}_{\mathfrak{J}',\pi}) &\leq 2\sqrt{\left(1 + \frac{p}{1 - Kp}\right) \frac{\epsilon^2}{(1 - Kp)\sigma^2} \cdot \frac{T}{K-1} + \frac{p\epsilon^2 T}{(1 - Kp)^2\sigma^2}} \\ &\leq 2\sqrt{\left(1 + \frac{p}{1 - Kp}\right) \frac{\epsilon^2}{(1 - Kp)\sigma^2} \cdot \frac{T}{K-1} + \frac{p\epsilon^2 T}{(1 - Kp)^2\sigma^2}}. \end{aligned}$$

Thus, from the definition of total variation,

$$\begin{aligned} & 1 + \mathbb{P}_{\mathfrak{J},\pi}(N_1 \geq T/2) - \mathbb{P}_{\mathfrak{J}',\pi}(N_1 < T/2) \\ &\geq 1 - d_{TV}(\mathbb{P}_{\mathfrak{J},\pi}, \mathbb{P}_{\mathfrak{J}',\pi}) \\ &\geq 1 - 2\sqrt{\left(1 + \frac{p}{1 - Kp}\right) \frac{\epsilon^2}{(1 - Kp)\sigma^2} \cdot \frac{T}{K-1} + \frac{p\epsilon^2 T}{(1 - Kp)^2\sigma^2}}. \end{aligned}$$

By picking $\epsilon = \min \left\{ \sqrt{\frac{(1-Kp)^2\sigma^2}{32pT}}, \sqrt{\frac{(K-1)(1-Kp)\sigma^2}{32\left(1+\frac{p}{1-Kp}\right)T}} \right\}$, we have $1 + \mathbb{P}_{\mathfrak{J},\pi}(N_1 \geq T/2) - \mathbb{P}_{\mathfrak{J}',\pi}(N_1 < T/2) \geq \frac{1}{2}$. Applying the above results to (B.54) gives,

$$\begin{aligned} & \mathbb{P}_{\mathfrak{J},\pi} \left(\text{Reg}_1^{\text{adv}}(T) \geq \frac{1}{4}\epsilon T \right) + \mathbb{P}_{\mathfrak{J}',\pi} \left(\text{Reg}_2^{\text{adv}'}(T) \geq \frac{1}{4}\epsilon T \right) \\ & \geq 1 + \mathbb{P}_{\mathfrak{J},\pi}(N_1 \geq T/2) - \mathbb{P}_{\mathfrak{J}',\pi}(N'_1 \geq T/2) - \mathbb{P}(W < 3T/4) - \mathbb{P}(W' < 3T/4) \\ & \geq \frac{1}{4}, \end{aligned}$$

where we use Proposition 8 (in Appendix B.2.2) to remove the terms involving W and W' .

This means either

$$\mathbb{P}_{\mathfrak{J},\pi} \left(\text{Reg}_1^{\text{adv}}(T) \geq \min \left\{ \sqrt{\frac{(1-Kp)^2\sigma^2 T}{32p}}, \sqrt{\frac{(K-1)(1-Kp)\sigma^2 T}{32\left(1+\frac{p}{1-Kp}\right)}} \right\} \right) \geq \frac{1}{8}$$

or

$$\mathbb{P}_{\mathfrak{J}',\pi} \left(\text{Reg}_2^{\text{adv}'}(T) \geq \min \left\{ \sqrt{\frac{(1-Kp)^2\sigma^2 T}{32p}}, \sqrt{\frac{(K-1)(1-Kp)\sigma^2 T}{32\left(1+\frac{p}{1-Kp}\right)}} \right\} \right) \geq \frac{1}{8},$$

which concludes the proof. □

B.2.2 Additional Propositions

Proposition 7. *Fix any $a \in (0, 1]$. We have*

$$\frac{x}{x + (1-a)x} \leq x + \frac{1-\sqrt{a}}{1+\sqrt{a}}, \quad \forall x \in (0, 1). \quad (\text{B.59})$$

Proof. It suffices to show, for any $a \in (0, 1]$, the function $f_a(x) := \frac{x}{x+(1-x)a} - x$ is upper bounded by $\frac{1-\sqrt{a}}{1+\sqrt{a}}$. This can be shown via a quick first-order test. At $x_{\max} = \frac{\sqrt{a}}{1+\sqrt{a}}$, the maximum of f_a is achieved, and $f_a(x_{\max}) = \frac{1-\sqrt{a}}{1+\sqrt{a}}$. □

Proposition 8. *Fix any $\sigma < \frac{1}{7}$. Pick T such that $T > 128 \log 8$, and ϵ such that $\frac{\epsilon}{1-Kp} \leq \frac{1}{4}$. Then $\mathbb{P}(W < \frac{3}{4}T) \leq \frac{1}{8}$ and $\mathbb{P}(W' < \frac{3}{4}T) \leq \frac{1}{8}$.*

Proof. Recall $W = \sum_{t=1}^T \mathbb{I}[\eta_t \in [-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}]]$. Since $\eta_t \in \mathcal{N}(0, \sigma^2)$, we have

$$\begin{aligned}
\mathbb{E} \left[\mathbb{I}[\eta_t \in [-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp}]] \right] &= \mathbb{P} \left(\eta_t \in \left[-\frac{1}{2}, \frac{1}{2} - \frac{\epsilon}{1-Kp} \right] \right) \\
&\geq \mathbb{P} \left(\eta_t \in \left[-\frac{1}{4}, \frac{1}{4} \right] \right) && \text{(since } \frac{\epsilon}{1-Kp} \leq \frac{1}{4} \text{)} \\
&= 1 - \mathbb{P} \left(|\eta_t| > \frac{1}{4} \right) \\
&= 1 - 2 \exp \left(-\frac{1}{16\sigma^2} \right) && \text{(since } \eta_t \text{ is } \sigma^2\text{-sub-Gaussian)} \\
&\geq \frac{7}{8}. && \text{(since } \sigma \leq \frac{1}{7} \text{)}
\end{aligned}$$

By Hoeffding's inequality,

$$\mathbb{P} \left(W < \frac{3}{4}T \right) \leq \mathbb{P} \left(W < \frac{7}{8}T - \sqrt{2T \log 8} \right) \leq \mathbb{P} \left(W < \mathbb{E}[W] - \sqrt{2T \log 8} \right) \leq \frac{1}{8}. \quad (\text{B.60})$$

□

Proposition 9. *For any distribution Q over problem instances and policy π , let $\mathbb{P}_{Q,\pi}$ be the probability of running π for T steps on a problem instance sampled from Q . For any problem instance \mathfrak{J} , let $\mathbb{P}_{\mathfrak{J},\pi}$ be the probability of running π for T steps on problem instance \mathfrak{J} . Then for any Q , π and event A and $u \in (0, 1)$, if $\mathbb{P}_{Q,\pi}(Q) \geq u$, then there exists $\mathfrak{J} \in \text{support}(Q)$, such that $\mathbb{P}_{\mathfrak{J},\pi}(A) \geq u$.*

Proof. For any event A ,

$$\mathbb{P}_{Q,\pi}(A) = \int_{\mathfrak{J} \in \text{support}(Q)} \mathbb{P}_{\mathfrak{J},\pi}(A) dQ(\mathfrak{J}). \quad (\text{B.61})$$

From above, it is clear that if $\mathbb{P}_{\mathfrak{J},\pi}(A) \leq u$ for all $\mathfrak{J} \in \text{support}(Q)$, then it is impossible to have $\mathbb{P}_{Q,\pi}(A) > u$.

□

Bibliography

- [AB09] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, pages 217–226, 2009.
- [ABM10] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.
- [AC16] Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016.
- [ACBDK15] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR, 2015.
- [ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- [ACBFS95] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.
- [ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
- [ACBG⁺17] Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.
- [ACBGM13] Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems*, pages 1610–1618, 2013.
- [AD16] Shipra Agrawal and Nikhil R Devanur. Linear contextual bandits with knapsacks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3458–3467, 2016.
- [AG12] Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.
- [AG13] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

- [Agr95a] Rajeev Agrawal. The continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33(6):1926–1951, 1995.
- [Agr95b] Rajeev Agrawal. Sample mean based index policies by $o(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.
- [AHT88] Rajeev Agrawal, MV Hedge, and Demosthenis Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.
- [AL17] Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.
- [AMS99] 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.
- [AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.
- [AO10] Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
- [AOS07] 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.
- [Ass83] Patrice Assouad. Plongements Lipschitziens dans \mathbb{R}^n . *Bulletin de la Société Mathématique de France*, 111:429–448, 1983.
- [Aue02] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- [AYBB⁺19] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.
- [AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.
- [B⁺65] Peter J Bickel et al. On some robust estimates of location. *The Annals of Mathematical Statistics*, 36(3):847–858, 1965.
- [B⁺15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends[®] in Machine Learning*, 8(3-4):231–357, 2015.

- [BB12] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.
- [BBBK11] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.
- [BBV11] Romain Benassi, Julien Bect, and Emmanuel Vazquez. Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion. *LION*, 5:176–190, 2011.
- [BCB⁺12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends[®] in Machine Learning*, 5(1):1–122, 2012.
- [BCBL13] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.
- [BFSO84] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.
- [BKS13] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.
- [BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.
- [BMS11] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.
- [BMSS11] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.
- [BS12] Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.
- [BSSM09] Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems*, pages 201–208, 2009.
- [BSY11] 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.

- [BT91] Dimitri P Bertsekas and John N Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.
- [Bul11] Adam D Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(Oct):2879–2904, 2011.
- [CBFH⁺97] Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P Helmbold, Robert E Schapire, and Manfred K Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.
- [CBL12] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.
- [CHM⁺15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.
- [CLQ17] Lijie Chen, Jian Li, and Mingda Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110, 2017.
- [CLRS11] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.
- [Cop09] 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.
- [CPV14] Emile Contal, Vianney Perchet, and Nicolas Vayatis. Gaussian process optimization with mutual information. In *International Conference on Machine Learning*, pages 253–261, 2014.
- [CS04] Imre Csiszár and Paul C Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
- [CSPD16] Roberto Calandra, André Seyfarth, Jan Peters, and Marc Peter Deisenroth. Bayesian optimization for learning gaits under uncertainty. *Annals of Mathematics and Artificial Intelligence*, 76(1-2):5–23, 2016.
- [CSS12] Michael Cwikel, Yoram Sagher, and Pavel Shvartsman. A new look at the John–Nirenberg and John–Strömberg theorems for BMO. *Journal of Functional Analysis*, 263(1):129–166, 2012.
- [CV95] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.
- [CWY13] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159. PMLR, 2013.

- [CZK19] Lin Chen, Mingrui Zhang, and Amin Karbasi. Projection-free bandit convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2047–2056. PMLR, 2019.
- [DDKP14] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467, 2014.
- [dFSZ12] Nando de Freitas, Alex Smola, and Masrour Zoghi. Exponential regret bounds for Gaussian process bandits with deterministic observations. In *International Conference on Machine Learning*, 2012.
- [DGS14] Travis Dick, Andras Gyorgy, and Csaba Szepesvari. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pages 512–520. PMLR, 2014.
- [DHK08] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.
- [DPG⁺14] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.
- [EDKM05] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Experts in a markov decision process. In *Advances in neural information processing systems*, pages 401–408, 2005.
- [EDKM09] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.
- [Fef79] Robert Fefferman. Bounded mean oscillation on the polydisk. *Annals of Mathematics*, 110(3):395–406, 1979.
- [FKM05] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.
- [FS97] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
- [Gar07] John B Garnett. Bounded mean oscillation. In *Bounded Analytic Functions*, pages 215–274. Springer, 2007.
- [GC11] Aurélien Garivier and Olivier Cappé. The KL–UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, pages 359–376, 2011.

- [GEW06] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.
- [Git79] John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.
- [GKX⁺14] Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *International Conference on Machine Learning*, pages 937–945, 2014.
- [GL16] Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems*, pages 1198–1206, 2016.
- [GNSA10] András György Gergely Neu, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Proceedings of the Twenty-Fourth Annual Conference on Neural Information Processing Systems*, 2010.
- [GSA14] Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.
- [H⁺16] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends[®] in Optimization*, 2(3-4):157–325, 2016.
- [Hei12] Juha Heinonen. *Lectures on analysis on metric spaces*. Springer Science & Business Media, 2012.
- [HHLB11] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer, 2011.
- [HKY17] Elad Hazan, Adam Klivans, and Yang Yuan. Hyperparameter optimization: A spectral approach. *arXiv preprint arXiv:1706.00764*, 2017.
- [HL14] Elad Hazan and Kfir Y Levy. Bandit convex optimization: Towards tight bounds. In *NIPS*, pages 784–792, 2014.
- [HW13] Eric Hall and Rebecca Willett. Dynamical models and tracking regret in online convex programming. In *International Conference on Machine Learning*, pages 579–587. PMLR, 2013.
- [ISSS19] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandr Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 202–219. IEEE, 2019.
- [JAZBJ18] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

- [JJL⁺19] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.
- [JJL⁺20] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.
- [Joh61] Fritz John. Rotation and strain. *Communications on Pure and Applied Mathematics*, 14(3):391–413, 1961.
- [JSW98] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.
- [JT16] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pages 240–248, 2016.
- [JYWJ20] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.
- [KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.
- [KBL⁺06] Oren S Klass, Ofer Biham, Moshe Levy, Ofer Malcai, and Sorin Solomon. The forbes 400 and the pareto wealth distribution. *Economics Letters*, 90(2):290–295, 2006.
- [KCG16] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.
- [KDO⁺16a] Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process optimisation with multi-fidelity evaluations. In *Proceedings of the 30th/International Conference on Advances in Neural Information Processing Systems (NIPS’30)*, 2016.
- [KDO⁺16b] Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabas Poczso. Multi-fidelity Gaussian process bandit optimisation. *arXiv preprint arXiv:1603.06288*, 2016.
- [KDSP17] Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabas Poczso. Multi-fidelity Bayesian optimisation with continuous approximations. *arXiv preprint arXiv:1703.06240*, 2017.
- [Kle05] Robert D Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.

- [KLSZ19] 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.
- [KNVM14] Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, pages 613–621, 2014.
- [KO11] Andreas Krause and Cheng S Ong. Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2011.
- [KP14] Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*, 2014.
- [KS06] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European Conference on Machine Learning*, pages 282–293. Springer, 2006.
- [KSH12a] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [KSH12b] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [KSU08] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *ACM Symposium on Theory of Computing*, pages 681–690. ACM, 2008.
- [KW97] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.
- [KWAS15] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.
- [LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web*, pages 661–670. ACM, 2010.

- [LeC99] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1999.
- [Ler13] Andrei K Lerner. The John–Nirenberg inequality with sharp constants. *Comptes Rendus Mathematique*, 351(11-12):463–466, 2013.
- [LJD⁺16] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. *The Journal of Machine Learning Research*, 2016.
- [LKOB17] Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained Bayesian optimization with noisy experiments. *arXiv preprint arXiv:1706.07094*, 2017.
- [LN18] Andrei K Lerner and Fedor Nazarov. Intuitive dyadic calculus: the basics. *Expositiones Mathematicae*, 2018.
- [LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- [LR85] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.
- [LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- [LW94] Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.
- [LWBS07] Daniel J Lizotte, Tao Wang, Michael H Bowling, and Dale Schuurmans. Automatic gait optimization with Gaussian process regression. In *IJCAI*, volume 7, pages 944–949, 2007.
- [LWHZ19] 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.
- [Mar] José Martell. 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.
- [MC14] Ruben Martinez-Cantin. Bayesopt: A bayesian optimization library for non-linear optimization, experimental design and bandits. *The Journal of Machine Learning Research*, 15(1):3735–3739, 2014.
- [MCP14] Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999, 2014.

- [MDY⁺17] Mustafa Mukadam, Jing Dong, Xinyan Yan, Frank Dellaert, and Byron Boots. Continuous-time Gaussian process motion planning via probabilistic inference. *arXiv preprint arXiv:1707.07383*, 2017.
- [MKS⁺15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- [MMS11] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Conference On Learning Theory*, pages 497–514, 2011.
- [MNSR17] Subhojyoti Mukherjee, K. P. Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Efficient-ucbv: An almost optimal algorithm using variance estimates. *arXiv preprint arXiv:1711.03591v1*, 2017.
- [MR11] Andrew McHutchon and Carl E Rasmussen. Gaussian process training with input noise. In *Advances in Neural Information Processing Systems*, 2011.
- [MS11] Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.
- [MT04] Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.
- [MY16] Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650, 2016.
- [NB13] Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *International Conference on Algorithmic Learning Theory*, pages 234–248. Springer, 2013.
- [Neu15] Gergely Neu. Explore no more: improved high-probability regret bounds for non-stochastic bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3168–3176, 2015.
- [NGSA10] Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 2*, pages 1804–1812, 2010.
- [NYW19] Chengzhuo Ni, Lin F Yang, and Mengdi Wang. Learning to control in metric space with optimal regret. In *57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 726–733. IEEE, 2019.

- [Pia14] Steven T Piantadosi. Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014.
- [PVG⁺11a] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [PVG⁺11b] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [Qui86] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
- [Rei82] P. Reimnitz. Asymptotic near admissibility and asymptotic near optimality by the “two armed bandit” problem. *Series Statistics*, 13(2):245–263, 1982.
- [RHW85] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
- [RM20] Aviv Rosenberg and Yishay Mansour. Adversarial stochastic shortest path. *arXiv preprint arXiv:2006.11561*, 2020.
- [Rob52] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
- [RRS15] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, 2015.
- [RS86] Richard Rochberg and Stephen Semmes. A decomposition theorem for BMO and applications. *Journal of functional analysis*, 67(2):228–263, 1986.
- [S⁺19] Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.
- [SB98] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.
- [SCY18] Jialin Song, Yuxin Chen, and Yisong Yue. A general framework for multi-fidelity Bayesian optimization with Gaussian processes. *arXiv preprint arXiv:1811.00755*, 2018.

- [SCY19] Jialin Song, Yuxin Chen, and Yisong Yue. A general framework for multi-fidelity bayesian optimization with gaussian processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3158–3167. PMLR, 2019.
- [Sha11] Ohad Shamir. A variant of Azuma’s inequality for martingales with sub-Gaussian tails. *arXiv preprint arXiv:1110.2392*, 2011.
- [Sha13] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24. PMLR, 2013.
- [SHK⁺14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [SHM⁺16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.
- [SKKS10a] Niranjana Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.
- [SKKS10b] Niranjana Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022, 2010.
- [SL17] Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759. Proceedings of Machine Learning Research, 2017.
- [SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.
- [Sli14] Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.
- [SM93] Elias M Stein and Timothy S Murphy. *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, volume 3. Princeton University Press, 1993.
- [Sno13] Jasper Roland Snoek. *Bayesian optimization and semiparametric models with applications to assistive technology*. PhD thesis, University of Toronto, 2013.

- [SS⁺11] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.
- [SS14] Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295, 2014.
- [SSA13] Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems*, pages 2004–2012, 2013.
- [SSW⁺16] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
- [SSZA14] Jasper Snoek, Kevin Swersky, Rich Zemel, and Ryan Adams. Input warping for bayesian optimization of non-stationary functions. In *International Conference on Machine Learning*, pages 1674–1682, 2014.
- [SV17] Leonid Slavin and Vasily Vasyunin. The John–Nirenberg constant of BMO^p , $1 \leq p \leq 2$. *St. Petersburg Mathematical Journal*, 28(2):181–196, 2017.
- [SWJ⁺20] 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, 2020.
- [SYKL18] Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, pages 8420–8429, 2018.
- [Tho33] 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.
- [TT98] Murad S Taqqu and Vadim Teverovsky. On estimating the intensity of long-range dependence in finite and infinite variance time series. *A practical guide to heavy tails: statistical techniques and applications*, 177:218, 1998.
- [TV15] Terence Tao and Van Vu. Random matrices: universality of local spectral statistics of non-hermitian matrices. *The Annals of Probability*, 43(2):782–874, 2015.
- [Utg89] Paul E Utgoff. Incremental induction of decision trees. *Machine learning*, 4(2):161–186, 1989.
- [VB07] Emmanuel Vazquez and Julien Bect. Convergence properties of the expected improvement algorithm. *arXiv preprint arXiv:0712.3744*, 2007.
- [VM05] Joannes Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *ECML*, volume 3720, pages 437–448. Springer, 2005.

- [Vu02] Van H Vu. Concentration of non-Lipschitz functions and applications. *Random Structures & Algorithms*, 20(3):262–316, 2002.
- [WDCW19] Yuanhao Wang, Kefan Dong, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. In *International Conference on Learning Representations*, 2019.
- [Whi88] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298, 1988.
- [WL18] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. *Proceedings of Machine Learning Research*, 75, 2018.
- [WMA⁺17] Tianyu Wang, Marco Morucci, M Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Flame: A fast large-scale almost matching exactly approach to causal inference. *arXiv preprint arXiv:1707.06315*, 2017.
- [WY19a] Nirandika Wanigasekara and Christina Yu. Nonparametric contextual bandits in an unknown metric space. In *Advances in Neural Information Processing Systems*, 2019.
- [WY19b] Nirandika Wanigasekara and Christina Lee Yu. Nonparametric contextual bandits in an unknown metric space. *ArXiv*, abs/1908.01228, 2019.
- [WY20] Tianyu Wang and Lin F. Yang. Episodic linear quadratic regulators with low-rank transitions. *arXiv:2011.01568*, 2020.
- [WYGR19] Tinayu Wang, Weicheng Ye, Dawei Geng, and Cynthia Rudin. Towards practical Lipschitz bandits. *arXiv preprint arXiv:1901.09277*, 2019.
- [WYGR20] Tianyu Wang, Weicheng Ye, Dawei Geng, and Cynthia Rudin. Towards practical lipschitz stochastic bandits. In *ACM-IMS Foundations of Data Science Conference*, 2020.
- [YW20] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Biography

Tianyu Wang is currently a PhD candidate at Computer Science Department of Duke University, advised by Prof. Cynthia Rudin. His research focuses on bandit learning, online learning, and reinforcement learning. Before coming to Duke, he obtained his B.Sc. in applied math and computer science from the Hone Kong University of Science and Technology (HKUST).