OPTIMAL LEARNING OF DEPLOYMENT AND SEARCH STRATEGIES FOR ROBOTIC TEAMS

By

Lai Wei

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Electrical and Computer Engineering – Doctor of Philosophy

2021

ABSTRACT

OPTIMAL LEARNING OF DEPLOYMENT AND SEARCH
STRATEGIES FOR ROBOTIC TEAMS

By

Lai Wei

In the problem of optimal learning, the dilemma of exploration and exploitation stems from the fact that gathering information and exploiting it are, in many cases, two mutually exclusive activities. The key to optimal learning is to strike a balance between exploration and exploitation. The Multi-Armed Bandit (MAB) problem is a prototypical example of such an explore-exploit tradeoff, in which a decision-maker sequentially allocates a single resource by repeatedly choosing one among a set of options that provide stochastic rewards. The MAB setup has been applied in many robotics problems such as foraging, surveillance, and target search, wherein the task of robots can be modeled as collecting stochastic rewards. The theoretical work of this dissertation is based on the MAB setup and three problem variations, namely heavy-tailed bandits, nonstationary bandits, and multi-player bandits, are studied. The first two variations capture two key features of stochastic feedback in complex and uncertain environments: heavy-tailed distributions and nonstationarity; while the last one addresses the problem of achieving coordination in uncertain environments. We design several algorithms that are robust to heavy-tailed distributions and nonstationary environments. Besides, two distributed policies that require no communication among agents are designed for the multi-player stochastic bandits in a piece-wise stationary environment.

The MAB problems provide a natural framework to study robotic search problems. The above variations of the MAB problems directly map to robotic search tasks in which a robot team searches for a target from a fixed set of view-points (arms). We further focus on the class of search problems involving the search of an unknown number of targets in a large or continuous space. We view the multi-target search problem as a hot-spots identification problem in which, instead of the global maximum of the field, all locations with a value greater than a threshold need to be identified. We consider a robot moving in 3D space with a downward-facing camera sensor. We model the robot's sensing output using a multi-fidelity Gaussian Process (GP) that systematically describes the sensing information available at different altitudes from the floor. Based on the sensing model, we design a novel algorithm that (i) addresses the coverage-accuracy tradeoff: sampling at a location farther from the floor provides a wider field of view but less accurate measurements, (ii) computes an occupancy map of the floor within a prescribed accuracy and quickly eliminates unoccupied regions from the search space, and (iii) travels efficiently to collect the required samples for target detection. We rigorously analyze the algorithm and establish formal guarantees on the target detection accuracy and the detection time.

An approach to extend the single robot search policy to multiple robots is to partition the environment into multiple regions such that workload is equitably distributed among all regions and then assign a robot to each region. The coverage control focuses on such equitable partitioning and the workload is equivalent to the so-called service demands in the coverage control literature. In particular, we study the adaptive coverage control problem, in which the demands of robotic service within the environment are modeled as a GP. To optimize the coverage of service demands in the environment, the team of robots aims to partition the environment and achieve a configuration that minimizes the coverage cost, which is a measure of the average distance of a service demand from the nearest robot. The robots need to address the explore-exploit tradeoff: to minimize coverage cost, they need to gather information about demands within the environment, whereas information gathering deviates them from maintaining a good coverage configuration. We propose an algorithm that schedules learning and coverage epochs such that its emphasis gradually shifts from exploration to exploitation while never fully ceasing to learn. Using a novel definition of coverage regret, we analyze the algorithm and characterizes its coverage performance over a finite time horizon.

ACKNOWLEDGEMENTS

Over the course of my PhD, I have received encouragement and support from many people. First and foremost, I would express my deep and most sincere gratitude to my advisor, Prof Vaibhav Srivastava. He guided me into the fascinating world of control and learning research and led my pathway to dissertation. I really enjoyed my research experience with him. He inspired me with his vision and expertise in every stage of my PhD study. I'll always remember his great mentorship, constant encouragement, and impeccable support.

I thank all my committee members, Prof Xiaobo Tan, Prof Ranjan Mukherjee, and Prof Shaunak Bopardikar for their great support and advice during my PhD study. Also, I thank them for teaching me adaptive control, nonlinear control and game theory. It had been a really nice experience to attend our robotics reading group and the National Robotics Initiative project meeting together.

I thank all my teachers at Michigan State University for their detailed explanation and quick responses to my questions. I particularly thank Prof Hassan Khalil for teaching me my first PhD course linear systems and control and ignite my interest in control theory. I'm also grateful to him for advising me before Dr. Srivastava joined MSU.

I thank every member of Prof Srivastava's, Prof Bopardikar's and Prof Tan's lab for their friendship, for their research presentations and for all our brainstorming sessions. Thank you Andrew McDonald for working together with me on coverage control. Thank you Pearce Reickert and Eric Gaskell for helping me design the robotic underwater target search experiment.

My stay at MSU is truly a fun experience. I'm blessed to make so many friends here. I thank Prof Ning Xi for taking me to MSU. Thank you Liangliang Chen for all the help when I first come to the U.S. and all the get-togethers at your home. Thank you Hongyang Shi for help me moving home. I'm so fortunate to meet with my girlfriend Kemeng Wang at MSU. Thank you so much for your constant love and encouragement. Thank you for being tolerant and supporting. Last, but not the least, I'd like to thank my mother and father for their unconditional love and endless support enabling me to pursue my dream.

The work reported in this dissertation was supported in part by NSF grant IIS-1734272 and ARO grant W911NF-18-1-0325.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

Decision-making in the face of uncertainty is one of the most fundamental problems in robotics research which requires the system to actively learn the environment while completing assigned tasks. Exploration versus exploitation tradeoff is a key challenge in such problems. Here, exploration means learning the environment to reduce the uncertainty while exploitation means taking the best actions according to the existing information. In applications such as robotic deployment and search where efficiency is of major concern, it is imperative to strike a good balance between exploration and exploitation. This encourages an investigation into algorithms that enable the robotic system to address this tradeoff in uncertain environments and achieve good finite-time performance.

The problems of interest in this dissertation include both theoretical investigations of the exploration versus exploitation tradeoff and its applications to robotic systems. The theoretical work is based on the Multi-Armed Bandit (MAB) setup [1] which is a classic mathematical formulation featuring the exploration versus exploitation tradeoff. With no prior information, a decision-maker sequentially chooses one among a set of stochastic arms (options) to achieve the maximum cumulative reward. An efficient adaptive arm selection rule keeps a good balance of learning the expected mean rewards and picking the empirically most profitable option. The high-level ideas embodied in the MAB problem extend naturally to many applications dealing with resource allocation in the face of uncertainty. Two prototypical examples in robotics research are target search and multi-robot deployment, and they are studied in this work. In a target search problem, to quickly and accurately locate the targets of interest, the autonomous vehicle needs to effectively learn the likelihood of target positions through sensing and spend more effort at locations that are more plausible to contain the target. The multi-robot deployment problem deals with the optimal allocation of a team of robots such that the robot team configuration matches with the demand of services within the environment. For example, more firefighting robots should be assigned to areas with a larger probability of wildfire breakouts to ensure a shorter response time.

In this dissertation, we study three variations of the MAB problem, namely heavy-tailed bandits, nonstationary bandits, and multi-player bandits. Though the classical stochastic MAB problem has a concise formulation and well-established theoretical guarantees, it fails to capture many key properties of stochastic processes in real-world problems. The heavy-tailed MAB relaxed the assumption that rewards from each arm are bounded or sub-Gaussian. This extension is motivated by applications such as social networks [2] and financial markets [3], wherein certain variables of interest exhibit heavy-tailed distributions. In the nonstationary MAB problem, in addition to being unknown, the reward distributions are assumed to be time-varying. This formulation characterizes the drift of physical processes in unknown dynamic working environments, e.g., the frequency of certain biological activities changes due to the different light conditioning in a day. The multi-player MAB problem involves maximizing the total group reward such that no two decision-makers select the same arm. This requires achieving a coordinated behavior of multiple decision-makers facing uncertainty. An important aspect of the multi-player MAB problems is the collision model, which means the reward is shared or no reward is received if multiple agents select the same option. This formulation arises naturally in cognitive radio [4–6], wherein a transceiver can intelligently detect and use vacant communication channels, and provides a rich modeling framework for other interesting domains such as animal and robotic foraging [7, 8], autonomous surveillance [9] and acoustic relay positioning for underwater communication [10].

A canonical example of the exploration-exploitation tradeoff in robotics applications is the target search problem. In many scenarios, the search task is to find the locations that emit large signals [11]. For example, the human body is relatively hot compared with its surroundings and emits more infrared radiations, which can be used to sense the existence of victims in a search and rescue scenario. The MAB problem provides a natural framework to study robotic search problems. In particular, the class of robotic search problems in which a robot team searches for a target from a set of view-points (arms), or monitors an environment from a set of viewpoints, directly map to the above variations of the MAB problems. We next focus on the class of search problems involving the search of an unknown number of targets in a large or continuous space that is not constrained to be observed from a small set of viewpoints.

We consider a robot moving in a 3D continuous environment to search for targets located on the 2D floor. For a given location of the robot, the sensors on the robot provide a score indicating the likelihood of the presence of a victim within the sensing footprint of the robot. We refer to these scores at all locations in the environment at a given altitude as the sensing field. The group of sensing fields at different altitudes from the floor is modeled a multi-fidelity Gaussian process (GP) [12] which is an auto-regressive model that captures the following fact: sensing at locations farther from the floor provides a wider field of view but less accurate measurements. The objective is to design a search policy that schedules the altitude and locations of the points at which the robot should collect samples such that the search time is minimized while ensuring a desired search accuracy.

An approach to extend the single robot search policy to *N* robots is to partition the environment into *N* regions such that some measure of "search load" is equitably distributed across all regions. The coverage control focuses on such equitable partitioning and the search load is typically referred to as serving demand in coverage control literature. In particular, the coverage control [13] concerns the deployment of a multi-agent team in order to meet "servicing" demands in an environment. The team of agents aims to minimize the coverage cost which is a function of the demand distribution, agent locations in the environment, and the allocation of the set of points within the environment to robots. In a standard coverage problem [13], the demand function is assumed to be known, while in this dissertation, we assume it to be unknown and model it as a realization of a GP. Similar to the MAB problem, the coverage problem with unknown demands exhibits the exploration versus exploitation tradeoff: learning the demands requires inefficient deployment with respect to the true demands.

1.1 Background and Literature Synopsis

The canonical stochastic MAB was historically proposed in [1] to model the clinical trials, where different treatments with unknown effects are viewed as stochastic arms. At each time step, the agent selects one arm from a set of options and receives a reward associated with it. The reward sequence at each arm is assumed to be an unknown i.i.d random process. The difficulty of such a problem lies in the exploration versus exploitation dilemma since the agent has to play the poorer arms to learn about their mean rewards. In fact, there exists an intrinsic tradeoff between choosing the most informative and seemingly the most rewarding alternative.

Robbins [14] formulated the objective of the stochastic MAB problem as minimizing the *regret*, that is, the loss in expected cumulative rewards caused by failing to select the best arm every time. In their seminal work, Lai and Robbins [15], established a logarithm *problem-dependent* asymptotic lower bound on the regret achieved by any policy. The lower bound being problem-dependent means it depends on the reward distribution at each arm. Using a simple heuristic called optimism in the face of uncertainty, a general method of constructing Upper Confidence Bound (UCB) rules for parametric families of reward distributions is also presented in [15], and the associated policy is shown to attain the logarithm lower bound. Several subsequent UCB-based algorithms [16, 17] with efficient finite-time performance have been proposed.

The simple formulation of the stochastic MAB limits its applications in many real-world problems. Bubeck et al. [18] relaxed the sub-Gaussian assumption by only assuming the rewards to have finite moments of order $1 + \epsilon$ for some $\epsilon > 0$. Their work allows the MAB model to be used in applications such as social networks [2] and financial markets [3] wherein certain variables of interest are inherently heavy-tailed.

The nonstationary MAB problem captures the dynamic aspect of the environment and has received some interest. In [19], the authors studied the bandit problem in which an adversary, rather than well-behaved stochastic arms, controls the payoffs. The performance of a policy is evaluated using the weak regret, which is the difference in the cumulated reward of a policy compared against the best single action policy. While being able to capture nonstationarity, the generality of the reward model in adversarial MAB makes the investigation of globally optimal policies very challenging. The nonstationary stochastic MAB can be viewed as a compromise between stationary stochastic MAB and adversarial MAB. It maintains the stochastic nature of the reward sequence while allowing some degree of nonstationarity in reward distributions. Instead of the weak regret analyzed in adversarial MAB, a strong notion of regret defined with respect to the best arm at each time step is studied in these problems. A broadly studied nonstationary problem is piecewise stationary MAB [20], wherein the reward distributions are piecewise stationary. A more general nonstationary problem is studied in [21], wherein the cumulative maximum variation in mean rewards is subject to a variation budget.

In some decision-making problems, multiple agents could get involved and the choice made by one agent could influence the selections of other agents. Most multi-player MAB studies deal with a stationary environment and the task is to maximize the total rewards collected by all the agents. As in the single-player case, the performance of the entire group of agents can be characterized using group regret, which is defined as the loss in expected total rewards caused by the agents failing to select the best set of arms every time. In [22], a lower bound on the group regret for a centralized policy is derived and algorithms that asymptotically achieve this lower bound are designed. Distributed multi-player MAB problem with no communication among players has been studied in [4, 5, 23–25]. In [26–28], distributed cooperative agents communicate through a communication network to improve their estimates of mean rewards and arm selections.

The MAB problem has been applied in many scientific and technological areas. For example, it is used for opportunistic spectrum access in communication networks, wherein multiple secondary users actively detect and use vacant channels [5, 29]. The arm models the availability of a channel and the reward from an arm is eliminated if it is selected by multiple users. In the MAB formulation of online learning for demand response [30, 31], an aggregator calls upon a subset of users (arms) who have an unknown response to the request to reduce their loads. Besides, contextual bandits are widely used in recommender systems [32, 33], wherein the acceptance of a recommendation corresponds to the rewards from an arm. The MAB setup has also been used in robotic research, which is one of the major topics in this dissertation.

In many robotic applications such as foraging, surveillance [7–9, 34] and acoustic relay positioning for underwater communication [10], the task of robots can be modeled to be collecting stochastic rewards [35, 36]. These rewards may correspond to, for example, the likelihood of an anomaly at a spatial location, the concentration of a certain type of algae in the ocean, the communication quality of a specific location, etc. Algorithms for the MAB problems have been extended to these problems. It needs to be noted that in robotic applications, motion is normally energy-consuming and time-demanding, while switch between arms comes with no cost in MAB models. By introducing block-allocation strategies, the exploration versus exploitation tradeoff can be balanced with a sufficiently small number of arm switches [37–39]. In [9], the robotic surveillance problem is studied in an environment that is abruptly changing due to the arrival of unknown spatial events. To solve the problem, a block-allocation strategy [20] is adapted to the piece-wise stationary MAB setting. Other algorithms that benefit path planning in robotic applications include the Deterministic Sequencing of Exploration and Exploitation (DSEE) algorithms [25, 40, 41] due to their deterministic and predictable structures.

Unlike the MAB problems in which rewards from different arms are commonly assumed to be independent, the feedbacks in robotic applications such as sensing information from different locations are usually correlated. GPs are powerful tools to capture spatially correlated information and they are widely used to characterizing spatiotemporal sensing fields [42, 43]. Sung *et al.* [11] study the hot-spot identification problem in an environment within the framework of GP MAB [38, 44]. GPs have also been used in robotic inspection and search. Hollinger *et al.* [45] study an inspection problem in which the robot needs to classify the underwater surface. They use a combination of GP-implicit surface modeling and sequential hypothesis testing to classify surfaces.

In autonomous robotic target search, the vehicle is required to quickly and accurately locate the targets of interest in an unknown and uncertain environment. Examples include victim search and rescue, mineral exploration, and tracking natural phenomena. There have been some efforts to address target search within the context of informative path planning [9, 11, 36, 45–58], which deals with path planning for the robots to maximize the utility of data collection. For example, Meera *et al.* [52] model target occupancy map as a GP and design a heuristic algorithm for target detection that handles tradeoffs among information gain, field coverage, sensor performance, and collision avoidance. In this dissertation, our multi-target search solution is inspired by successive elimination ideas from MAB research in [59, 60]. The robot sequentially collected new sensing information and remove regions unlikely to contain the target from the search task. The proposed algorithm combines informative path planning with Bayesian confidence interval estimation and enables the robot to efficiently collect information and concentrate measurements in promising areas.

The coverage control [13] is another interesting topic in robotics. It arises naturally in multi-robot systems when a team of agents is assigned to deploy themselves over an environment according to a particular demand function, which specifies the degree to which a robot is needed at each location. The objective is to minimize the coverage cost which is determined by the demand distribution, robot locations, and task allocations. Example applications of coverage control range from autonomous wildfire fighting, to smart agriculture, to ecological surveying, to environmental cleaning. Classical coverage control problem [13, 61–63] assumes the demand function is known, while recent works have focused on the adaptive coverage problem, in which agents are not assumed to have knowledge of demand function a priori. In [64–69], the demand function defined on the working environment is modeled as a realization of a GP which can be learned by taking samples at different locations. The exploration versus exploitation dilemma in adaptive coverage control is due to the conflict between collecting samples to refine demand estimation and maintaining a good configuration to reduce the coverage cost. As with the MAB setup, adaptive coverage also solves a stochastic optimization problem, though its task is more complicated than selecting the most rewarding arm. In the same spirit as the DSEE algorithm for MAB problems, the adaptive coverage policy proposed in this dissertation deterministically shifts its emphasis from exploration to exploitation.

1.2 Contribution and Organization

In this section, the organization of the chapters in this dissertation is outlined and the contributions in each chapter are discussed in detail.

Chapter 2. We first review the stationary stochastic MAB problem and corresponding concepts such as regret, worst-case regret, and lower bounds on regret. We then study the heavy-tailed Bandits in which the sub-Gaussian assumption of reward distributions is relaxed. Instead, the reward distributions admit moments of order $1+\epsilon$ for some $\epsilon > 0$, similarly as in [18]. We modify the MOSS [70] algorithm for the sub-Gaussian reward distribution by using a saturated empirical mean to design a new algorithm called Robust MOSS. By analyzing Robust MOSS, we show that it achieves worst-case regret matching with the lower bound while maintaining a distribution-dependent logarithm regret. To the best of our knowledge, Robust MOSS is the first algorithm to achieve order-optimal worst-case regret for heavy-tailed bandits. This is the major contribution of this chapter. Numerical illustrations are provided to verify the robustness of the proposed algorithm against heavy-tailed rewards. We close the chapter with comments and bibliographic remarks on the classic MAB problem.

Chapter 3. In this chapter, we study a special nonstationary stochastic MAB problem called piece-wise stationary bandits. We assume the mean rewards from arms switch to unknown values at unknown time instants and the reward distribution remains stationary between consecutive switches.

The main contribution of this chapter is the design of two generic algorithms, namely, the Limited Memory DSEE (LM-DSEE) and the Sliding-Window UCB# (SW-UCB#). LM-DSEE inherits the structure of the DSEE algorithm [25, 40] for the stationary bandit and comprises interleaving blocks of exploration and exploitation. In the exploitation epochs, an arm is selected based on the most recent exploration. This avoids a large bias in reward estimation in a nonstationary environment. And, SW-UCB# is a modification of the SW-UCB algorithm from [20] that relaxes the assumption of knowing horizon length. We rigorously show these algorithms incur sublinear regret, i.e., the time average of the regret asymptotically converges to zero. A comparison of both algorithms is made and discussed based on the simulation results.

Chapter 4. We study the multi-player stochastic bandits in a piece-wise stationary environment. We consider a collision model in which a player receives a reward at an arm if it is the only player to select the arm. This problem features achieving coordination in an uncertain and nonstationary environment.

The contribution of this chapter is the design of two novel distributed algorithms that require no communication between agents, namely Round-Robin SW-UCB# (RR-SW-UCB#) and the Sliding-Window Distributed Learning with Prioritization (SW-DLP). For both algorithms, it is shown that the group regret is upper bounded by a sublinear function of time even with the collision model, i.e., both algorithms achieve coordination while efficiently learning a nonstationary environment.

Chapter 5. In this chapter, we study a more general nonstationary stochastic MAB problem proposed in [21] in which the cumulative maximum variation in mean rewards is restricted to a variation budget. There is no restriction on how do reward distributions change, for example, they may change abruptly like the piece-wise stationary bandits, or they may drift slowly between subsequent abrupt changes. The performance of a policy is measured by comparing its cumulative expected rewards with that of an oracle that selects the arm with the maximum mean reward at each time and it is characterized using the worst-case regret, which is the regret for the worst choice of reward distribution sequences that satisfies the variation budget. We extend UCB-based policies with three different approaches, namely, periodic resetting, sliding observation window, and discount factor, and show that they achieve order-optimal worst-case performance. We also relax the sub-Gaussian assumption on reward distributions and develop robust versions of the proposed policies that can handle heavy-tailed reward distributions and maintain their performance guarantees.

The major contributions of this work are threefold. First, we extend MOSS [70] to design Resetting MOSS (R-MOSS) and Sliding-Window MOSS (SW-MOSS). Also, we show Discounted UCB (D-UCB) [20] can be tuned to solve the problem. Secondly, with rigorous analysis, we show that R-MOSS and SW-MOSS achieve the exact order-optimal worst-case performance and D-UCB is near-optimal. Finally, we relax the bounded or sub-Gaussian assumption on the rewards required by these algorithms and design policies robust to heavy-tailed rewards. We show the theoretical guarantees on the worst-case regret can be maintained by the robust policies.

Chapter 6. We consider a scenario in which an autonomous vehicle equipped with a downward-facing camera operates in a 3D environment and is tasked with searching for an unknown number of stationary targets on the 2D floor of the environment. The key challenge is to design a search policy that minimizes the search time while ensuring a high target detection accuracy. We model the sensing field using a multi-fidelity GP that systematically describes the sensing information available at different altitudes from the floor. Based on the sensing model, we design a novel algorithm called Expedited Multi-Target Search (EMTS) that (i) addresses the coverage-accuracy tradeoff: sampling at a location farther from the floor provides a wider field of view but less accurate measurements, (ii) computes an occupancy map of the floor within a prescribed accuracy and quickly eliminates unoccupied regions from the search space, and (iii) travels efficiently to collect the required samples for target detection. We rigorously analyze the algorithm and establish formal guarantees on the target detection accuracy and the expected detection time. We illustrate the algorithm using a simulated multi-target search scenario.

The primary contribution of this chapter is the extension of the classical informative path planning approach for the single-fidelity GP to the multi-fidelity GP setting. This novel extension allowed for jointly planning for sampling locations and associated fidelity levels, and thus, addresses the fidelity-coverage tradeoff for expedited target search. The EMTS algorithm is proposed and illustrated in an underwater victim search scenario using the Unmanned Underwater Vehicle Simulator. The algorithm is analyzed in terms of its accuracy and expected detection time.

Chapter 7. We study the problem of distributed multi-robot coverage over an unknown, nonuniform sensory field, which is a deployment problem with uncertain demands. Modeling the sensory field as a realization of a GP and using Bayesian techniques, we devise a policy that aims to balance the tradeoff between learning the sensory function and covering the environment. We propose an adaptive coverage algorithm called Deterministic Sequencing of Learning and Coverage (DSLC) that schedules learning and coverage epochs such that its emphasis gradually shifts from exploration to exploitation while never fully ceasing to learn. Using a novel definition of coverage regret which characterizes the overall coverage performance of a multi-robot team over a time horizon T, we analyze DSLC to provide an upper bound on expected cumulative coverage regret. Finally, we illustrate the empirical performance of the algorithm through simulations of the coverage task over an unknown distribution of wildfires.

The most important contribution of this chapter is the definition of coverage regret which enables the finite-time analysis of the online estimation and coverage algorithm. Existing works evaluate the algorithm performance by assuming the coverage algorithm attains global optimality and comparing the performance with a globally optimal result. Since the coverage problem itself is NP-Hard, this assumption is too strong, especially with distributed deployment being considered. The coverage regret is defined with respect to locally optimal solutions which relaxes the assumption of achieving a globally optimal coverage and perfectly characterizes the convergence property of a policy.

Chapter 8. In this chapter, we conclude the dissertation with a summary of contributions and future directions. For the MAB problems addressed in this work, we discuss their potential applications in robotic patrolling. For the robotic target search and adaptive coverage control, the problem generalizations and possible solutions are illustrated in detail.

CHAPTER 2

STATIONARY STOCHASTIC BANDITS

In a stationary stochastic MAB problem, an agent chooses an arm φ_t from the set of K arms $\{1, \ldots, K\}$ at each time $t \in \{1, \ldots, T\}$ and receives the associated random reward $X_t^{\varphi_t}$. The reward at each arm k is drawn from an unknown probability distribution f_k with unknown mean μ_k. The problem being stationary mean the reward distribution for each arm does not change with time. Normally, it is assumed that the reward distributions are sub-Gaussian.

Definition 2.1 (Sub-Gaussian reward)**.** *For any arm $k \in \{1, \ldots, K\}$, the probability distribution f_k is $1/2$ sub-Gaussian, i.e., if $X \sim f_k$,*

$$\forall \lambda \in \mathbb{R} : \mathbb{E}\left[\exp(\lambda(X - \mu))\right] \leq \exp\left(\frac{\lambda^2}{8}\right).$$

The main example for sub-Gaussia rewards are random rewards with bounded support $[0, 1]$ which is commonly used in MAB literature.

The objective for the stochastic MAB problem is to maximize the expected value of the *cumulative reward* $\sum_{t=1}^{T} X_t^{\varphi_t}$. We assume that φ_t is selected based upon past observations $\{X_s^{\varphi_s}, \varphi_s\}_{s=1}^{t-1}$ following some policy ρ. Specifically, ρ determines the conditional distribution

$$\mathbb{P}^{\rho}\left(\varphi_t = k \mid \{X_s^{\varphi_s}, \varphi_s\}_{s=1}^{t-1}\right)$$

at each time $t \in \{1, \ldots, T - 1\}$. If $\mathbb{P}^{\rho}(\cdot)$ takes binary values, we call ρ deterministic; otherwise, it is called stochastic.

Let the maximum mean reward among all arms be μ^*. We use $\Delta_k = \mu^* - \mu_k$ to measure the suboptimality of arm k. For a policy ρ, to maximize the expected cumulative reward $\mathbb{E}[S_T]$ is equivalent to minimize the *regret* defined by

$$R_T^{\rho} := \mathbb{E}^{\rho}\left[\sum_{t=1}^{T}\left(\mu^* - X_t^{\varphi_t}\right)\right] = \sum_{k=1}^{K} \Delta_k \mathbb{E}^{\rho}[n_k(T)],$$

where $n_k(T)$ is the total number of times the arm k has been chosen until time T, and the second expectation is taken over different realization of arm selections. It needs to be noted that R_T^{ρ} can be viewed as the difference between the expected cumulative reward obtained by selecting the arm with the maximum mean reward μ^* and selecting arms $\varphi_1, \dots, \varphi_T$.

2.1 Lower Bounds on Regrets

The objective of regret minimization was originally formulated by Robbins [14]. It was established later a logarithm *problem-dependent* asymptotic lower bound on the number of times a suboptimal arm is selected by a uniformly good policy in [15, 71]. Here, a policy ρ being uniformly good means for any possible set of reward distributions $\{f_1, \dots, f_K\}$,

$$\mathbb{E}\,[R_T^\rho] = o(T^a) \quad \text{for every } a > 0,$$

which means $\lim_{T\to\infty} \mathbb{E}\,[R_T^\rho]/T^a = 0$ for any $a > 0$.

Lemma 2.1 (Lai and Robbins' Lower bound [15, 71])**.** *Suppose there is a unique best arm with reward distribution f^* and an uniformly good policy ρ is applied. For any suboptimal arm k and every $\epsilon > 0$,*

$$\lim_{T\to\infty} \mathbb{P}\left(n_k(T) \geq \frac{1-\epsilon}{D_{\mathrm{KL}}(f_k||f^*)}\right) = 1,$$

where D_{KL} denote the Kullback-Leibler divergence of two distributions. Hence,

$$\liminf_{T\to\infty} \frac{\mathbb{E}\,[n_k(T)]}{\log T} \geq \frac{1}{D_{\mathrm{KL}}(f_k||f^*)}.$$

The above result indicates a suboptimal arm needs to be selected at least logarithm number of times, resulting in the logarithm lower bound for a stochastic MAB problem. It can also be seen that regret R_T^ρ is implicitly determined by reward distributions in $\{f_1, \dots, f_K\}$ as well as policy ρ. So, R_T^ρ is also called *distribution-dependent* regret. In contrast, the *distribution-independent* regret, also known as worst-case regret, is defined by taking the maximum over at all possible reward distribution combinations.

Definition 2.2 (Worst-case Regret)**.** *The worst-case regret is the regret for the worst possible choice of reward distributions and it can be expressed as*

$${}^{\text{worst}}R_T^\rho = \sup_{\{f_1,\dots,f_K\}} R_T^\rho.$$

The regret associated with the policy that minimizes the above worst-case regret is called *minimax regret*. According to [72], the minimax regret also has a lower bound $1/20\sqrt{KT}$. This result is about finite-time performance and can be derived by selecting a set of reward distributions that present challenges to the allocation policy. Consider a scenario in which there is a unique best arm and all other arms have identical mean rewards such that the gap between optimal and suboptimal mean rewards is Δ. For such a problem, it has been shown in [73] that for any policy ρ,

$$R_T^\rho \geq C_1 \frac{K}{\Delta} \ln\left(\frac{T\Delta^2}{K}\right) + C_2 \frac{K}{\Delta}, \tag{2.1}$$

where C_1 and C_2 are some positive constants. It needs to be noted that for $\Delta = \sqrt{K/T}$, the above lower bound becomes $C_2\sqrt{KT}$, which matches with the lower bound $1/20\sqrt{KT}$.

2.2 Upper Confidence Bound Strategies

The family of Upper Confidence Bound (UCB) strategies uses the principle called optimism in the face of uncertainty. At each time slot, a UCB index which is a statistical index composed of both mean reward estimate and the associated uncertainty measure is computed at each arm, and the arm with the maximum UCB is picked. Within the family of UCB algorithms, two state-of-the-art algorithms for the stationary stochastic MAB problems are UCB1 [16] and MOSS [70]. With arm k being sampled $n_k(t)$ of times before t, $\hat{\mu}_{k,n_k(t)}$ is the associated empirical mean. Then, UCB1 computes the UCB index for each arm k at time t as

$$g_{k,t}^{\text{UCB1}} = \hat{\mu}_{k,n_k(t)} + \sqrt{\frac{2\ln t}{n_k(t)}}.$$

The finite-time performance guarantee of UCB1, which is stronger than asymptotic property, has been proved in [16].

Lemma 2.2 (Regret Upper Bound for UCB1)**.** *For the stationary stochastic MAB problem, if the reward distributions have bounded support* $[0, 1]$*, the regret of UCB1 after any T satisfies*

$$R_T^{\text{UCB1}} \leq 8 \sum_{k:\Delta_k>0} \frac{\ln T}{\Delta_k} + \left(1 + \frac{\pi^2}{3}\right) \sum_{k=1}^{K} \Delta_k.$$

Notice that the above upper bound matches with the order of Lai and Robbin's logarithm lower bound in Lemma 2.1. As is shown in [70], the worst-case regret of UCB1 can be derived by selecting values for Δ_k to maximize the upper bound, resulting in

$$^{\text{worst}}R_T^{\text{UCB1}} \leq 10\sqrt{(K-1)T(\ln T)}.$$

Comparing this result with the lower bound on the minimax regret $1/20\sqrt{KT}$ there exists an extra factor $\sqrt{\ln T}$. This issue has been resolved by the algorithm called Minimax Optimal Strategy in the Stochastic case (MOSS) [70], which is the first algorithm that enjoys both logarithm distribution-dependent and order-optimal distribution-independent bound. With prior knowledge of horizon length T, and the UCB index for MOSS is expressed as

$$g_{k,t}^{\text{MOSS}} = \hat{\mu}_{k,n_k(t)} + \sqrt{\frac{\max\left(\ln\left(\frac{T}{Kn_k(t)}\right), 0\right)}{n_k(t)}}.$$

We now recall the worst-case regret upper bound for MOSS.

Lemma 2.3 (Worst-case regret upper bound for MOSS [70])**.** *For the stationary stochastic MAB problem, the worst-case regret of the MOSS algorithm satisfies*

$$^{\text{worst}}R_T^{\text{MOSS}} \leq 49\sqrt{KT}.$$

2.3 Heavy-tailed Stochastic MAB

This section is a slightly modified version of our published work on heavy-tailed bandits, and it is reproduced here with the permission of the copyright holder[1].

The rewards being bounded or sub-Gaussian is a common assumption that gives the sample mean an exponential convergence and simplifies the MAB problem. However, in many applications, such as social networks [2] and financial markets [3], the rewards are heavy-tailed. Bubeck et al. [18] relax the sub-Gaussian assumption by only assuming the rewards to have finite moments of order $1+\epsilon$ for some $\epsilon \in (0,1]$. They present the robust UCB algorithm and show that it attains a

[1] ©2018 IEEE. Reprinted with permission from [74].

logarithmic distribution-depend regret upper bound on the regret that is within a constant factor of the lower bound in the heavy-tailed setting. However, the solutions provided in [18] are not able to provably achieve an order optimal worst-case regret. Specifically, the factor of optimality is a poly-logarithmic function of time-horizon.

The heavy-tailed stochastic MAB problem studied in this section is the stochastic MAB problem with the following assumptions.

Assumption 2.1. *Let X be a random reward drawn from any arm $k \in \{1, \dots, K\}$. There exists a constant $u \in \mathbb{R}_{>0}$ such that $\mathbb{E}\left[|X|^{1+\epsilon}\right] \leq u^{1+\epsilon}$ for some $\epsilon \in (0, 1]$.*

Assumption 2.2. *Parameters T, K, u and ϵ are known.*

We now recall the lower bound on the minimax regret for the heavy tailed bandit problem derived in [18].

Theorem 2.4 ([18, Th. 2])**.** *For any fixed time horizon T and the stochastic MAB problem under Assumptions 2.1 and 2.2 with $u = 1$, the worst-case regret for a uniformly good policy ρ satisfies*

$$^{\text{worst}}R_T^{\rho} \geq 0.01 K^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}.$$

Remark 2.1. *Since R_T scales with u, the lower bound for heavy tail bandit is $\Omega\big(uK^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}\big)$. This lower bound also indicates that within a finite horizon T, it is almost impossible to differentiate the optimal arm from arm k, if $\Delta_k \in O\big(u(K/T)^{\frac{\epsilon}{1+\epsilon}}\big)$. As a special case, rewards with bounded support $[0, 1]$ correspond to $\epsilon = 1$ and $u = 1$. Then, the lower bound $\Omega(\sqrt{KT})$ matching with the regret upper bound is achieved by MOSS.*

2.3.1 A Robust Minimax Policy: Robust MOSS

In Robust MOSS, to deal with the heavy-tailed reward distribution, we replace the empirical mean with a saturated empirical mean. Although the saturated empirical mean is a biased estimator, it has better convergence properties. The formal definition is given later in this section. We construct a novel UCB index to evaluate the arms, and at each time slot, the arm with the maximum UCB index is picked. Let $n_k(t)$ be the number of times that arm k has been selected until time $t-1$. At time t, let $\hat{\mu}^k_{n_k(t)}$ be the saturated empirical mean reward computed from the $n_k(t)$ samples at arm k. Robust MOSS initializes by selecting each arm once and subsequently, at each time t, selects the arm that maximizes the following UCB

$$g^k_{n_k(t)} = \hat{\mu}^k_{n_k(t)} + (1+\eta)c_{n_k(t)},$$

where $\eta > 0$ is an appropriate constant, $c_{n_k(t)} = u \times \left[\phi(n_k(t))\right]^{\frac{\epsilon}{1+\epsilon}}$ and

$$\phi(n) = \frac{\ln_+\left(\frac{T}{Kn}\right)}{n},$$

where $\ln_+(x) := \max(\ln x, 1)$. Note that both $\phi(n)$ and c_n are monotonically decreasing in n.

The robust saturated empirical mean is similar to the truncated empirical mean used in [18], which is employed to extend UCB1 to achieve logarithm distribution-dependent regret for the heavy-tailed MAB problem. Let $\{X_i\}_{i\in\{1,\dots,m\}}$ be a sequence of i.i.d. random variables with mean μ and $\mathbb{E}\left[|X_i|^{1+\epsilon}\right] \leq u^{1+\epsilon}$, where $u > 0$. Pick $a > 1$ and let $h(m) = a^{\lfloor \log_a(m) \rfloor + 1}$ such that $h(m) \geq m$. Define the saturation point B_m by

$$B_m := u \times \left[\phi\big(h(m)\big)\right]^{-\frac{1}{1+\epsilon}}.$$

Then, the saturated empirical mean estimator is defined by

$$\hat{\mu}_m := \frac{1}{m}\sum_{i=1}^{m} \mathrm{sat}(X_i, B_m), \tag{2.2}$$

where $\mathrm{sat}(X_i, B_m) := \mathrm{sign}(X_i)\min\left\{|X_i|, B_m\right\}$.

Define $d_i := \mathrm{sat}(X_i, B_m) - \mathbb{E}\left[\mathrm{sat}(X_i, B_m)\right]$ which has the following properties.

Lemma 2.5. *For any $i \in \{1,\dots,m\}$, d_i satisfies (i) $|d_i| \leq 2B_m$ (ii) $\mathbb{E}\left[d_i^2\right] \leq u^{1+\epsilon}B_m^{1-\epsilon}$.*

Proof. Property (i) follows immediately from the definition of d_i, and property (ii) follows from

$$\mathbb{E}\left[d_i^2\right] \leq \mathbb{E}\left[\,\mathrm{sat}^2(X_i, B_m)\right] \leq \mathbb{E}\left[|X_i|^{1+\epsilon}\, B_m^{1-\epsilon}\right].$$

□

The following lemma examines the estimator bias and provides an upper bound on the error of the saturated empirical mean.

Lemma 2.6 (Bias of saturated empirical mean)**.** *For an i.i.d. sequence of random variables $\{X_i\}_{i\in\{1,\dots,m\}}$ such that $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}\left[X_i^{1+\epsilon}\right] \leq u^{1+\epsilon}$, the saturated empirical mean* (2.2) *satisfies*

$$\left|\hat{\mu}_m - \mu - \frac{1}{m}\sum_{i=1}^{m} d_i\right| \leq \frac{u^{1+\epsilon}}{B_m^{\epsilon}}.$$

Proof. Since $\mu = \mathbb{E}\left[X_i\left(\mathbf{1}_{\{|X_i|\leq B_m\}} + \mathbf{1}_{\{|X_i|>B_m\}}\right)\right]$, the error of estimator $\hat{\mu}_m$ satisfies

$$\hat{\mu}_m - \mu = \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{sat}(X_i, B_m) - \mu\right) = \frac{1}{m}\sum_{i=1}^{m} d_i + \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}\left[\mathrm{sat}(X_i, B_m)\right] - \mu\right),$$

where the second term is the bias of $\hat{\mu}_m$. We now compute an upper bound on the bias.

$$\left|\mathbb{E}\left[\mathrm{sat}(X_i, B_m)\right] - \mu\right| \leq \mathbb{E}\left[|X_i|\,\mathbf{1}_{\{|X_i|>B_m\}}\right] \leq \mathbb{E}\left[\frac{|X_i|^{1+\epsilon}}{(B_m)^{\epsilon}}\right] \leq \frac{u^{1+\epsilon}}{(B_m)^{\epsilon}},$$

which concludes the proof. □

2.3.2 Analysis of Robust MOSS

In this section, we analyze Robust MOSS to provide both distribution-free and distribution-dependent regret bounds.To derive the concentration property of saturated empirical mean, we use a maximal Bennett type inequality as shown in Lemma 2.7.

Lemma 2.7 (Maximal Bennett's inequality [75])**.** *Let $\{X_i\}_{i\in\{1,\dots,n\}}$ be a sequence of bounded random variables with support $[-B, B]$, where $B \geq 0$. Suppose that $\mathbb{E}[X_i|X_1,\dots,X_{i-1}] = \mu_i$ and $\mathrm{Var}[X_i|X_1,\dots,X_{i-1}] \leq v$. Let $S_m = \sum_{i=1}^{m}(X_i - \mu_i)$ for any $m \in \{1,\dots,n\}$. Then, for any $\delta \geq 0$*

$$\mathbb{P}\left(\exists m \in \{1,\dots,n\} : S_m \geq \delta\right) \leq \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right),$$

$$\mathbb{P}\left(\exists m \in \{1,\dots,n\} : S_m \leq -\delta\right) \leq \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right),$$

where $\psi(x) = (1 + 1/x)\ln(1 + x) - 1$.

Remark 2.2. *For $x \in (0, \infty)$, function $\psi(x)$ is monotonically increasing in x.*

Now, we establish an upper bound on the probability that the UCB underestimates the mean at arm k by an amount x.

Lemma 2.8. *For any arm $k \in \{1, \ldots, K\}$ and any $t \in \{K+1, \ldots, T\}$ and $x > 0$, if $\eta\psi(2\eta/a) \geq 2a$, the probability of event $\left\{g^k_{n_k(t)} \leq \mu_k - x\right\}$ is no greater than*

$$\frac{K}{T}\frac{a}{\ln(a)}\Gamma\left(\frac{1}{\epsilon}+2\right)\left(\frac{\psi\left(2\eta/a\right)}{2a}\frac{x}{u}\right)^{-\frac{1+\epsilon}{\epsilon}}.$$

Proof. It follows from Lemma 2.6 that

$$\begin{aligned}\mathbb{P}\left(g^k_{n_k(t)} \leq \mu_k - x\right) \leq & \mathbb{P}\left(\exists m \in \{1, \ldots, T\} : \hat{\mu}^k_m + (1+\eta)c_m \leq \mu_k - x\right)\\ \leq & \mathbb{P}\left(\exists m \in \{1, \ldots, T\} : \sum_{i=1}^{m}\frac{d^k_i}{m} \leq \frac{u^{1+\epsilon}}{B^{\epsilon}_m} - (1+\eta)c_m - x\right)\\ \leq & \mathbb{P}\left(\exists m \in \{1, \ldots, T\} : \frac{1}{m}\sum_{i=1}^{m} d^k_i \leq -x - \eta c_m\right),\end{aligned}$$

where d^k_i is defined similarly to d_i for i.i.d. reward sequence at arm k and the last inequality is due to

$$\frac{u^{1+\epsilon}}{B^{\epsilon}_m} = u\left[\phi\left(h(m)\right)\right]^{\frac{\epsilon}{1+\epsilon}} \leq u\left[\phi(m)\right]^{\frac{\epsilon}{1+\epsilon}} = c_m. \tag{2.3}$$

Recall $a > 1$. We apply a peeling argument [76, Sec 2.2] with geometric grid $a^s \leq m < a^{s+1}$ over time interval $\{1, \ldots, T\}$. Since c_m is monotonically decreasing with m,

$$\begin{aligned}&\mathbb{P}\left(\exists m \in \{1, \ldots, T\} : \frac{1}{m}\sum_{i=1}^{m} d^k_i \leq -x - \eta c_m\right)\\ \leq & \sum_{s\geq 0}\mathbb{P}\left(\exists m \in [a^s, a^{s+1}) : \sum_{i=1}^{m} d^k_i \leq -a^s\left(x + \eta c_{a^{s+1}}\right)\right).\end{aligned}$$

Also notice that $B_m = B_{a^s}$ for all $m \in [a^s, a^{s+1})$. Then with properties in Lemma 2.5, we apply

Lemma 2.7 to get

$$
\begin{aligned}
&\sum_{s\geq 0}\mathbb{P}\left(\exists m\in[a^s,a^{s+1}) : \sum_{i=1}^{m} d_i^k \leq -a^s\left(x+\eta c_{a^{s+1}}\right)\right)\\
\leq&\sum_{s\geq 0}\exp\left(-\frac{a^s\left(x+\eta c_{a^{s+1}}\right)}{2B_{a^s}}\psi\left(\frac{2B_{a^s}\left(x+\eta c_{a^{s+1}}\right)}{au^{1+\epsilon}B_{a^s}^{1-\epsilon}}\right)\right)\\
&\left(\ \psi(x) \text{ is monotonically increasing}\right)\\
\leq&\sum_{s\geq 0}\exp\left(-\frac{a^s\left(x+\eta c_{a^{s+1}}\right)}{2B_{a^s}}\psi\left(\frac{2\eta B_{a^s}^{\epsilon}c_{a^{s+1}}}{au^{1+\epsilon}}\right)\right)\\
&\left(\text{plug in } c_{a^{s+1}},\ B_{a^s} \text{ and use } h(a^s)=a^{s+1}\right)\\
=&\sum_{s\geq 1}\exp\left(-a^s\left(\frac{x}{B_{a^{s-1}}}+\eta\phi(a^s)\right)\frac{\psi\left(2\eta/a\right)}{2a}\right).
\end{aligned}
$$

Plugging in $\phi(a^s)$, with $\eta\psi(2\eta/a)\geq 2a$ and $\ln_+(y)\geq\ln(y)$, we have

$$
\sum_{s\geq 1}\exp\left(-a^s\left(\frac{x}{B_{a^{s-1}}}+\eta\phi(a^s)\right)\frac{\psi\left(2\eta/a\right)}{2a}\right)\leq\sum_{s\geq 1}\exp\left(-a^s\frac{x}{B_{a^{s-1}}}\frac{\psi\left(2\eta/a\right)}{2a}\right)\frac{K}{T}a^s.
$$

Let $b=x\psi\left(2\eta/a\right)/(2au)$. Since $B_{a^{s-1}}\leq ua^{\frac{s}{1+\epsilon}}$, we have

$$
\begin{aligned}
\sum_{s\geq 1}\exp\left(-a^s\frac{x}{B_{a^{s-1}}}\frac{\psi\left(2\eta/a\right)}{2a}\right)\frac{K}{T}a^s\leq&\frac{K}{T}\sum_{s\geq 1}a^s\exp\left(-ba^{\frac{\epsilon s}{1+\epsilon}}\right)\\
\leq&\frac{K}{T}\int_1^{+\infty}a^y\exp\left(-ba^{\frac{(y-1)\epsilon}{1+\epsilon}}\right)dy\\
=&\frac{K}{T}a\int_0^{+\infty}a^y\exp\left(-ba^{\frac{y\epsilon}{1+\epsilon}}\right)dy\\
&\left(\text{where we set } z=ba^{\frac{y\epsilon}{1+\epsilon}}\right)\\
=&\frac{K}{T}\frac{a}{\ln(a)}\frac{1+\epsilon}{\epsilon}b^{-\frac{1+\epsilon}{\epsilon}}\int_b^{+\infty}z^{\frac{1+\epsilon}{\epsilon}-1}\exp\left(-z\right)dz\\
\leq&\frac{K}{T}\frac{a}{\ln(a)}\Gamma\left(\frac{1}{\epsilon}+2\right)b^{-\frac{1+\epsilon}{\epsilon}},
\end{aligned}
$$

which concludes the proof. □

The following is a straightforward corollary of Lemma 2.8.

Corollary 2.9. *For any arm $k\in\{1,\dots,K\}$ and any $t\in\{K+1,\dots,T\}$ and $x>0$, if $\eta\psi(2\eta/a)\geq 2a$, the probability of event $\left\{g_{n_k(t)}^k-2(1+\eta)c_{n_k(t)}\geq\mu_k+x\right\}$ shares the same bound in Lemma 2.8.*

The distribution-free upper bound for Robust MOSS, which is the main result for the paper, is presented in this section. We show that the algorithm achieves order optimal worst-case regret.

Theorem 2.10. *For the heavy-tailed stochastic MAB problem with K arms and time horizon T, if η and a are selected such that $\eta\psi(2\eta/a) \geq 2a$, then Robust MOSS satisfies*

$$^{\text{worst}}R_T^{\text{Robust MOSS}} \leq CuK^{\frac{\epsilon}{1+\epsilon}}(T/e)^{\frac{1}{1+\epsilon}} + 2uK,$$

where

$$\begin{aligned} C =& \Gamma\left(1/\epsilon+2\right)\left[a/(6+3\eta)\right]^{\frac{1}{\epsilon}}\left[3/\psi\left(6+3\eta\right)\right]^{\frac{1+\epsilon}{\epsilon}} \\ &+ \epsilon\Gamma\left(1/\epsilon+2\right)\left(6+3\eta\right)^{-\frac{1}{\epsilon}}\left[6a/\psi(2\eta/a)\right]^{\frac{1+\epsilon}{\epsilon}}a/\ln(a) + \left(6+3\eta\right)\left[e+(1+\epsilon)e^{\frac{-\epsilon}{1+\epsilon}}\right]. \end{aligned}$$

Remark 2.3. *Parameter a and η as inputs to Robust MOSS can be selected by minimizing the leading constant C in the upper bound on the regret in Theorem 2.10. We have found that selecting a slightly larger than* 1 *and selecting smallest η that satisfies $\eta\psi(2\eta/a) \geq 2a$ yields good performance.*

Proof. Since both the UCB and the regret scales with u defined in Assumption 2.1, to simplify the expressions, we assume $u = 1$. Also notice that Assumption 2.1 indicates $\left|\mu_k\right| \leq u$, so $\Delta_k \leq 2$ for any $k \in \{1, \ldots, K\}$. In the following, any terms with superscript or subscript “$*$" and “k" are with respect to the best and the k-th arm, respectively. The proof is divided into 4 steps.

Step 1: We follow a decoupling technique inspired by the proof of regret upper bound in MOSS [70]. Take the set of δ-bad arms as $\mathcal{B}_\delta$ as

$$\mathcal{B}_\delta := \{k \in \{1, \ldots, K\} \mid \Delta_k > \delta\}, \tag{2.4}$$

where we assign $\delta = \left(6+3\eta\right)\left(eK/T\right)^{\frac{\epsilon}{1+\epsilon}}$. Thus,

$$\begin{aligned} R_T &\leq T\delta + \sum_{t=1}^{K}\Delta_k + \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\}\left(\Delta_{\varphi_t}-\delta\right)\right] \\ &\leq T\delta + 2K + \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\}\left(\Delta_{\varphi_t}-\delta\right)\right]. \end{aligned} \tag{2.5}$$

Furthermore, we make the following decomposition

$$\begin{aligned}\sum_{t=K+1}^{T} \mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\} \left(\Delta_{\varphi_t} - \delta\right) = &\sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} \leq \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\} \left(\Delta_{\varphi_t} - \delta\right) \\ &+ \sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} > \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\} \left(\Delta_{\varphi_t} - \delta\right).\end{aligned} \tag{2.6}$$

Notice that the first summand (2.6) describes regret from underestimating optimal arm $*$. For the second summand, since $g^{\varphi_t}_{n_{\varphi_t}(t)} \geq g^*_{n^*(t)}$ and $\mu^* = \mu_{\varphi_t} + \Delta_{\varphi_t}$,

$$\begin{aligned}&\sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} > \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\} \left(\Delta_{\varphi_t} - \delta\right) \\ \leq &\sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^{\varphi_t}_{n_{\varphi_t}(t)} > \mu_{\varphi_t} + \frac{2\Delta_{\varphi_t}}{3}\right\}\Delta_{\varphi_t} \\ = &\sum_{k \in \mathcal{B}_\delta} \sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t = k, g^k_{n_k(t)} > \mu_k + \frac{2\Delta_k}{3}\right\}\Delta_k,\end{aligned} \tag{2.7}$$

which characterizes the regret caused by overestimating δ-bad arms.

Step 2: In this step, we bound the expectation of (2.6). When event $\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} \leq \mu^* - \Delta_{\varphi_t}/3\right\}$ happens, we know

$$\Delta_\varphi \leq 3\mu^* - 3g^*_{n^*(t)} \text{ and } g^*_{n^*(t)} < \mu^* - \frac{\delta}{3}.$$

Thus, we get

$$\mathbf{1}\left\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} \leq \mu^* - \frac{\Delta_{\varphi_t}}{3}\right\}(\Delta_{\varphi_t} - \delta) \leq \mathbf{1}\left\{g^*_{n^*(t)} < \mu^* - \frac{\delta}{3}\right\} \times \left(3\mu^* - 3g^*_{n^*(t)} - \delta\right) := Y_t$$

Since Y_t is a positive random variable, its expected value can be computed involving only its cumulative density function:

$$\begin{aligned}\mathbb{E}\left[Y_t\right] = \int_0^{+\infty} \mathbb{P}\left(Y_t > x\right) dx &\leq \int_0^{+\infty} \mathbb{P}\left(3\mu^* - 3g^*_{n^*(t)} - \delta > x\right) dx \\ &= \int_\delta^{+\infty} \mathbb{P}\left(\mu^* - g^*_{n^*(t)} > \frac{x}{3}\right) dx.\end{aligned}$$

Then we apply Lemma 2.8 at optimal arm $*$ to get

$$\mathbb{E}\left[Y_t\right] \leq \frac{KC_1}{T} \int_\delta^{+\infty} \frac{1}{\epsilon} x^{-\frac{1+\epsilon}{\epsilon}} dx = \frac{KC_1}{T\delta^{\frac{1}{\epsilon}}}$$

where $C_1 = \epsilon\Gamma\left(1/\epsilon + 2\right)\left[6a/\psi(2\eta/a)\right]^{\frac{1+\epsilon}{\epsilon}} a/\ln(a)$. We conclude this step by

$$\mathbb{E}\Big[\sum_{t=K+1}^{T} \mathbf{1}\Big\{\varphi_t \in \mathcal{B}_\delta, g^*_{n^*(t)} \le \mu^* - \frac{\Delta_{\varphi_t}}{3}\Big\}\left(\Delta_{\varphi_t} - \delta\right)\Big] \le \sum_{t=K+1}^{T} Y_t \le C_1 K \delta^{-\frac{1}{\epsilon}}.$$

Step 3: In this step, we bound the expectation of (2.7). For each arm $k \in \mathcal{B}_\delta$,

$$\begin{aligned}
&\sum_{t=K+1}^{T} \mathbf{1}\Big\{\varphi_t = k, g^k_{n_k(t)} \ge \mu_k + \frac{2\Delta_k}{3}\Big\} \\
&= \sum_{t=K+1}^{T} \sum_{m=1}^{t-K} \mathbf{1}\left\{\varphi_t = k, n_k(t) = m\right\} \mathbf{1}\left\{g^k_m \ge \mu_k + \frac{2\Delta_k}{3}\right\} \\
&= \sum_{m=1}^{T-K} \mathbf{1}\left\{g^k_m \ge \mu_k + \frac{2\Delta_k}{3}\right\} \sum_{t=m+K}^{T} \mathbf{1}\left\{\varphi_t = k, n_k(t) = m\right\} \\
&\le \sum_{m=1}^{T} \mathbf{1}\left\{g^k_m \ge \mu_k + \frac{2\Delta_k}{3}\right\} \\
&\le \sum_{m=1}^{T} \mathbf{1}\Big\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \ge \frac{2\Delta_k}{3} - (2+\eta)c_m\Big\},
\end{aligned} \tag{2.8}$$

where in the last inequality we apply Lemma 2.6 and use the fact that $u^{1+\epsilon}/B^\epsilon_m \le c_m$ in (2.3). We set

$$l_k = \left\lceil \left(\frac{6+3\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}} \ln\left(\frac{T}{K}\left(\frac{\Delta_k}{6+3\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\right\rceil.$$

With $\Delta_k \ge \delta$, we get l_k is no less than

$$\left(\frac{6+3\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}} \ln\left(\frac{T}{K}\left(\frac{\delta}{6+3\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right) = \left(\frac{6+3\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}}.$$

Furthermore, since c_m is monotonically decreasing with m, for $m \ge l_k$,

$$c_m \le c_{l_k} \le \left[\frac{\ln_+\left(\frac{T}{K}\left(\frac{\Delta_k}{6+3\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right)}{l_k}\right]^{\frac{\epsilon}{1+\epsilon}} \le \frac{\Delta_k}{6+3\eta}. \tag{2.9}$$

With this result and $l_k \ge 1$, we continue from (2.8) to get

$$\begin{aligned}
\mathbb{E}\Big[\sum_{m=1}^{T} \mathbf{1}\Big\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \ge \frac{2\Delta_k}{3} - (2+\eta)c_m\Big\}\Big] \le& l_k - 1 + \sum_{m=l_k}^{T} \mathbb{P}\Big\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \ge \frac{2\Delta_k}{3} - (2+\eta)c_m\Big\} \\
\le& l_k - 1 + \sum_{m=l_k}^{T} \mathbb{P}\Big\{\frac{1}{m}\sum_{i=1}^{m} d^k_i \ge \frac{\Delta_k}{3}\Big\}
\end{aligned} \tag{2.10}$$

Therefore by using Lemma 2.7 together with statement (ii) from Lemma 2.5, we get

$$\sum_{m=l_k}^{T} \mathbb{P}\left\{\frac{1}{m}\sum_{i=1}^{m} d_i^k \geq \frac{\Delta_k}{3}\right\} \leq \sum_{m=l_k}^{T} \exp\left(-\frac{m\Delta_k}{3B_m}\psi\left(B_m^\epsilon \Delta_k\right)\right) \leq \sum_{m=l_k}^{T} \exp\left(-\frac{m\Delta_k}{3B_m}\psi\left(6+3\eta\right)\right),$$

where the last step is due to that $\psi(x)$ is monotonically increasing and $B_m^\epsilon \Delta_k \geq (6{+}3\eta)B_m^\epsilon c_m \geq 6{+}3\eta$ from (2.9) and (2.3). Since $B_m = \phi\big(h(m)\big)^{-\frac{1}{1+\epsilon}} \leq \phi(am)^{-\frac{1}{1+\epsilon}} \leq (am)^{\frac{1}{1+\epsilon}}$, we have

$$\begin{aligned}\sum_{m=l_k}^{T} \exp\left(-\frac{m\Delta_k}{3B_m}\psi\left(6+3\eta\right)\right) &\leq \sum_{m=1}^{T} \exp\left(-m^{\frac{\epsilon}{1+\epsilon}} a^{-\frac{1}{1+\epsilon}}\psi\left(6+3\eta\right)\frac{\Delta_k}{3}\right). \\ &\leq \int_0^{+\infty} \exp\left(-\beta y^{\frac{\epsilon}{1+\epsilon}}\right) dy\end{aligned}$$

where we set $\beta = a^{-\frac{1}{1+\epsilon}}\psi\left(6+3\eta\right)\Delta_k/3$. Taking $z = \beta y^{\frac{\epsilon}{1+\epsilon}}$, we obtain

$$\int_0^{+\infty} \exp\left(-\beta y^{\frac{\epsilon}{1+\epsilon}}\right) dy = \frac{1+\epsilon}{\epsilon}\beta^{-\frac{1+\epsilon}{\epsilon}} \int_0^{+\infty} z^{\frac{1+\epsilon}{\epsilon}-1}\exp\left(-z\right) dy = \Gamma\left(\frac{1}{\epsilon}+2\right)\beta^{-\frac{1+\epsilon}{\epsilon}}.$$

Plugging it into (2.10),

$$\mathbb{E}\left[\sum_{m=1}^{T}\mathbf{1}\left\{\frac{1}{m}\sum_{i=1}^{m} d_i^k \geq \frac{2\Delta_k}{3} - (2+\eta)c_m\right\}\right] \leq C_2\Delta_k^{-\frac{1+\epsilon}{\epsilon}} + C_3\Delta_k^{-\frac{1+\epsilon}{\epsilon}}\ln\left(\frac{T}{KC_3}\Delta_k^{\frac{1+\epsilon}{\epsilon}}\right)$$

where $C_2 = \Gamma\left(1/\epsilon+2\right)a^{\frac{1}{\epsilon}}\left[3/\psi\left(6+3\eta\right)\right]^{\frac{1+\epsilon}{\epsilon}}$ and $C_3 = \left(6+3\eta\right)^{\frac{1+\epsilon}{\epsilon}}$. Putting it together with $\Delta_k \geq \delta$ for all $k \in \mathcal{B}_\delta$, the expectation of (2.7) is no greater than

$$\sum_{k\in\mathcal{B}_\delta} C_2\Delta_k^{-\frac{1}{\epsilon}} + C_3\Delta_k^{-\frac{1}{\epsilon}}\ln\left(\frac{T}{KC_3}\Delta_k^{\frac{1+\epsilon}{\epsilon}}\right) \leq C_2K\delta^{-\frac{1}{\epsilon}} + (1+\epsilon)e^{\frac{-\epsilon}{1+\epsilon}}C_3K\delta^{-\frac{1}{\epsilon}},$$

where we use the fact that $x^{-\frac{1}{\epsilon}}\ln\left(Tx^{\frac{1+\epsilon}{\epsilon}}/(KC_3)\right)$ takes its maximum at $x = \delta\exp(\epsilon^2/(1+\epsilon))$.

Step 4: Plugging the results in step 2 and step 3 into (2.5),

$$^{\text{worst}}R_T^{\text{Robust MOSS}} \leq T\delta + \left[C_1 + C_2 + (1+\epsilon)e^{\frac{-\epsilon}{1+\epsilon}}C_3\right]K\delta^{-\frac{1}{\epsilon}} + 2K.$$

Straightforward calculation concludes the proof. □

We now show that robust MOSS also preserves a logarithmic upper bound on the distribution-dependent regret.

Theorem 2.11. *For the heavy-tailed stochastic MAB problem with K arms and time horizon T, if $\eta\psi(2\eta/a) \geq 2a$, the regret R_T for Robust MOSS is no greater than*

$$\sum_{k:\Delta_k>0} \left(\frac{u^{1+\epsilon}}{\Delta_k}\right)^{\frac{1}{\epsilon}} \left[C_1 \ln\left(\frac{T}{KC_1}\left(\frac{\Delta_k}{u}\right)^{\frac{1+\epsilon}{\epsilon}}\right) + C_2 K\right] + \Delta_k,$$

where $C_1 = (4+4\eta)^{\frac{1+\epsilon}{\epsilon}}$ and $C_2 = \max\left(eC_1, 2\Gamma(1/\epsilon+2)\left(8a/\psi(2\eta/a)\right)^{\frac{1+\epsilon}{\epsilon}} a/\ln(a)\right)$.

Proof. Let $\delta = (4+4\eta)\left(eK/T\right)^{\frac{\epsilon}{1+\epsilon}}$ and define $\mathcal{B}_\delta$ the same as (2.4). Since $\Delta_k \leq \delta$ for all $k \notin \mathcal{B}_\delta$, the regret satisfies

$$\begin{aligned} R_T^{\text{Robust MOSS}} &\leq \sum_{k\notin\mathcal{B}_\delta} T\Delta_k + \sum_{t=1}^{T} \mathbf{1}\{\varphi_t \in \mathcal{B}_\delta\}\Delta_{\varphi_t} \\ &\leq \sum_{k\notin\mathcal{B}_\delta} eK\left(\frac{4+4\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}}\Delta_k + \sum_{k\in\mathcal{B}_\delta}\sum_{t=1}^{T} \mathbf{1}\{\varphi_t = k\}\Delta_k. \end{aligned} \tag{2.11}$$

Pick arbitrary $l_k \in \mathbb{Z}_+$, thus

$$\begin{aligned} \sum_{t=1}^{T} \mathbf{1}\{\varphi_t = k\} &\leq l_k + \sum_{t=K+1}^{T} \mathbf{1}\left\{\varphi_t = k, n_k(t) \geq l_k\right\} \\ &\leq l_k + \sum_{t=K+1}^{T} \mathbf{1}\left\{g^k_{n_k(t)} \geq g^*_{n^*(t)}, n_k(t) \geq l_k\right\}. \end{aligned}$$

Observe that $g^k_{n_k(t)} \geq g^*_{n^*(t)}$ implies at least one of the following is true

$$g^*_{n^*(t)} \leq \mu^* - \Delta_k/4, \tag{2.12}$$

$$g^k_{n_k(t)} \geq \mu_k + \Delta_k/4 + 2(1+\eta)c_{n_k(t)}, \tag{2.13}$$

$$(1+\eta)c_{n_k(t)} > \Delta_k/4. \tag{2.14}$$

We select

$$l_k = \left\lceil \left(\frac{4+4\eta}{\Delta_k}\right)^{\frac{1+\epsilon}{\epsilon}} \ln\left(\frac{T}{K}\left(\frac{\Delta_k}{4+4\eta}\right)^{\frac{1+\epsilon}{\epsilon}}\right)\right\rceil.$$

Similarly as (2.9), $n_k(t) \geq l_k$ indicates $c_{n_k(t)} \leq \Delta_k/(4+4\eta)$, so (2.14) is false. Then we apply Lemma 2.8 and Corollary 2.9,

$$\mathbb{P}\left\{g^k_{n_k(t)} \geq g^*_{n^*(t)}, n_k(t) \geq l_k\right\} \leq \mathbb{P}\left(\text{(2.12) or (2.13) is true }\right) \leq \frac{C_2' K}{T}\Delta_k^{-\frac{1+\epsilon}{\epsilon}},$$

where $C_2' = 2\Gamma\left(1/\epsilon + 2\right)\left(8a/\psi(2\eta/a)\right)^{\frac{1+\epsilon}{\epsilon}} a/\ln(a)$. Substituting it into (2.11),

$$R_T^{\text{Robust MOSS}} \leq \sum_{k \notin \mathcal{B}_\delta} \frac{eC_1 K}{\Delta_k^{\frac{1}{\epsilon}}} + \sum_{k \in \mathcal{B}_\delta} \left[\frac{C_1}{\Delta_k^{\frac{1}{\epsilon}}} \ln\left(\frac{T}{KC_1} \Delta_k^{\frac{1+\epsilon}{\epsilon}} \right) + \frac{C_2' K}{\Delta_k^{\frac{1}{\epsilon}}} + \Delta_k \right].$$

Considering the scaling factor u, the proof can be concluded with easy computation. □

2.3.3 Numerical Illustration of Robust MOSS

In this section, we compare Robust MOSS with MOSS and Robust UCB (with truncated empirical mean or Catoni's estimator) [18] in a 3-armed heavy-tailed bandit setting. The mean rewards are $\mu_1 = -0.3$, $\mu_2 = 0$ and $\mu_3 = 0.3$ and sampling at each arm k returns a random reward equals to μ_k added by sampling noise ν, where$|\nu|$ is a generalized Pareto random variable and the sign of ν has equal probability to be positive and negative. The PDF of reward at arm k is

$$f_k(x) = \frac{1}{2\sigma}\left(1 + \frac{\xi|x - \mu_k|}{\sigma}\right)^{-\frac{1}{\xi}-1} \quad \text{for } x \in (-\infty, +\infty),$$

where we select $\xi = 0.33$ and $\sigma = 0.32$. Thus, for a random reward X from any arm, we know $\mathbb{E}\left[X^2\right] \leq 1$, which means $\epsilon = 1$ and $u = 1$. We select parameters $a = 1.1$ and $\eta = 2.2$ for Robust MOSS so that condition $\eta\psi(2\eta/a) \geq 2a$ is met.

Figure.2.1 shows the mean regret together with quantiles of regret distribution as a function of time, which are computed using 200 simulations of each policy. On each graph, the bold curve is the emprical mean regret while light shaded and dark shaded regions correspond respectively to upper 5% and lower 95% quantile cumulative regrets. The simulation result shows that there is a chance MOSS loses stability in heavy-tailed MAB and suffers linear regret while other algorithms work consistently and maintain sub-linear regrets. Robust MOSS slightly outperforms Robust UCB in this specific problem.

2.4 Summary

We reviewed the stationary stochastic MAB problem and concepts including regret and regret lower bound. Specially, we studied the heavy-tailed bandit problem and proposed the Robust

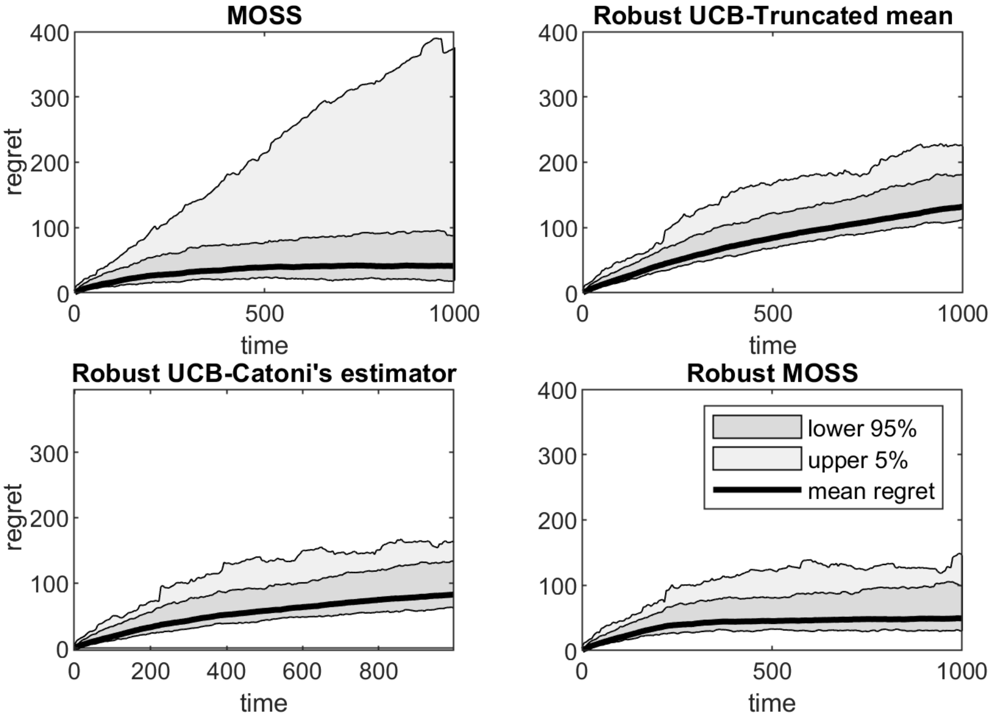


Figure 2.1: Comparison of Robust MOSS with MOSS and other Robust UCB algorithms.

MOSS algorithm. We evaluate it by deriving upper bounds on the associated distribution-free and distribution-dependent regrets. Our analysis shows that Robust MOSS achieves order optimal performance in both scenarios. It can be noticed that the saturated mean estimator centers at zero so that the algorithm is not translation invariant. Exploration of translation invariant robust mean estimator in this context remains an open problem.

2.5 Bibliographic Remarks

Since the seminal work by Lai and Robbins [15], several subsequent works design simpler algorithms by assuming that the rewards are bounded or more generally, sub-Gaussian. By using Kullback-Leibler(KL) divergence-based uncertainty estimates, Garivier and Cappé [17] designed KL-UCB and proved that it strictly dominates UCB1 [16], which uses Hoeffding inequality-based uncertainty estimates. Aside from the nonBayesian policies mentioned above, Bayesian strategies have also been proved to be effective for MAB problems. The Bayes-UCB algorithm by Kaufmann et al. [77] is the first Bayesian algorithm proved to be asymptotic optimal. Thompson sampling [1], proposed in 1933, has long been shown to perform very well in practice. Very recently, asymptotic and finite-time performance guarantees that are very close to the optimal have been proved for Thompson sampling [78, 79]. Other effective policies include ϵ-greedy [16]and deterministic sequencing of exploration and exploitation [25, 40, 41]. However, both these classes of algorithms require the knowledge of a lower bound on the minimum gap in mean rewards.

In the context of minimizing the worst-case regret, Ménard and Garivier [80] adapted the MOSS algorithm with KL divergence-based uncertainty estimates and proposed a minimax algorithm kl-UCB^{++} that improves the algorithm performance. It also needs to be noticed that MOSS requires knowing horizon length, so it is not an anytime algorithm. Degenne and Perchet [81] extend MOSS to an any-time version called MOSS-anytime that can adapt to different horizon lengths.

CHAPTER 3

PIECE-WISE STATIONARY STOCHASTIC BANDITS

In nonstationary stochastic MAB problem, the reward sequence $\{X_t^k\}_{t=1}^T$ at each arm $k \in \{1, \ldots, K\}$ is composed of independent samples from time-varying reward distributions $\{f_t^k\}_{t=1}^T$. Piece-wise stationary MAB is a special type of the nonstationary bandit problem in which f_t^k switches at unknown time instants referred as *breakpoints*. Between consecutive breakpoints, f_t^k remains the same for any $k \in \{1, \ldots, K\}$. In this chapter, we assume each f_t^k has bounded support $[0, 1]$, and the total number of breakpoints until time T is $\Upsilon_T \in O(T^\nu)$, where $\nu \in [0, 1)$ and is known a priori.

Similarly as the stationary Stochastic MAB problem, the decision-maker's objective is to select arms $\varphi_t, \ldots, \varphi_T$ that maximizes the expected of cumulative reward $\mathbb{E}\left[\sum_{t=1}^T X_t^{\varphi_t}\right]$. Let the time-varying mean reward associated with arm k be μ_t^k at time $t \in \{1, \ldots, T\}$. Then, for the nonstationary MAB problem, the regret for a policy ρ can be defined by

$$R_T^\rho := \sum_{t=1}^T \mu_t^* - \mathbb{E}\left[\sum_{t=1}^T X_t^{\varphi_t}\right] = \mathbb{E}^\rho\left[\sum_{t=1}^T \mu_t^* - \mu_t^{\varphi_t}\right], \tag{3.1}$$

where $\mu_t^* = \max_{k\in\{1,\ldots,K\}} \mu_t^k$ and the expectation is with respect to different realization of $\{\varphi_t\}_{t=1}^T$ that depends on obtained rewards through policy ρ.

This chapter is a slightly modified version of our published work on piece-wise stationary stochastic bandits, and it is reproduced here with the permission of the copyright holder[1]. In the following sections, two generic algorithms, namely Limited-Memory DSEE (LM-DSEE) algorithm and the Sliding-Window UCB# (SW-UCB#) algorithm, are presented and analyzed. These algorithms require parameters to be tuned based on environment characteristics.

3.1 Preliminaries

We first recall two MAB policies since algorithms proposed in this chapter are developed based upon them. The first algorithm is Deterministic Sequencing of Exploration and Exploitation (DSEE) [40].

[1] ©2018 IEEE. Reprinted with permission from [82].

It divides the set of natural numbers $\mathbb{N}$ into interleaving blocks of exploration and exploitation. In the exploration block all arms are played in a round-robin fashion, while in the exploitation block, the arm with the maximum statistical mean reward is played. For an appropriately defined $w \in \mathbb{R}_{>0}$, the DSEE algorithm at time t exploits if the number of exploration steps until time $t-1$ are greater than or equal to $K\lceil w \log t \rceil$, otherwise it starts a new exploration block. Vakili *et al.* showed that the DSEE algorithm achieves efficient performance. It should be noted that tuning w requires knowledge of a lower bound on the gap between the mean reward from the best arm and the second-best arm. This requirement can be relaxed at the cost of degraded performance.

The second policy is Sliding-Window UCB (SW-UCB) [20]. It is a variation of UCB1 [16] that intends to solve the piecewise-stationary bandits. A sliding observation window is used to erase the outdated sampling history, and the UCB index is computed within it. Since the size of the sliding observation window in SW-UCB depends on the horizon length, it requires knowledge of the horizon length of the problem. The SW-UCB# proposed in this chapter intends to relax this assumption and enable the policy to adapt to different horizon lengths.

3.2 The LM-DSEE Algorithm

The LM-DSEE algorithm comprises interleaving blocks of exploration and exploitation. In the n-th exploration epoch, each arm is sampled $L(n) = \lceil \gamma \ln(n^{\varrho} l b) \rceil$ number of times. In the n-th exploitation epoch, the arm with the highest sample mean in the n-th exploration epoch is sampled $\lceil a n^{\varrho} l \rceil - K L(n)$ times. Here, the parameters ϱ, γ, a, b, and l are tuned based on the environment characteristics (see Algorithm 1 for details). In the following, we will set a and b to unity for the purposes of analysis.

The LM-DSEE algorithm is similar in spirit to the DSEE algorithms [25, 40], wherein the length of the exploitation epoch increases exponentially with the epoch number, and all the data collected in the previous exploration epochs are used to estimate mean rewards. However, in a non-stationary environment using all the rewards from the previous exploration epochs may lead to a heavily biased estimate of mean rewards. Furthermore, an exponentially increasing exploitation

Algorithm 1: The LM-DSEE Algorithm

```
Input   : $\nu \in [0, 1), \Delta_{\min} \in (0, 1), T \in \mathbb{N}, a \in \mathbb{R}_{>0}, b \in (0, 1]$;
Set     : $\gamma \geq \frac{2}{\Delta_{\min}^2}$, $al \in \{K\lceil \gamma \ln lb \rceil, \ldots, +\infty\}$, and $\varrho = \frac{1-\nu}{1+\nu}$;
Output  : sequence of arm selection;
% Initialization:
Set batch index $n \leftarrow 1$ and $t \leftarrow 1$;
while $t \leq T$ do
    % Exploration
    for $k \in \{1, \ldots, K\}$ do
        Pick arm $k$, $L(n) \leftarrow \lceil \gamma \ln(n^{\varrho} lb) \rceil$ times ;
        collect rewards $\{X_i^k(n)\}_{i \in \{1, \ldots, L(n)\}}$ ;
        compute sample mean $\bar{\mu}_k^{\text{epch}}(n) \leftarrow \frac{1}{L(n)} \sum_{i=1}^{L(n)} X_i^k(n)$;
    % Exploitation
    Select the best arm $\varphi_n^{\text{epch}} = \arg\max_{k \in \{1, \ldots, K\}} \bar{\mu}_k^{\text{epch}}(n)$ ;
    Pick arm $\varphi_n^{\text{epch}}$, $\lceil an^{\varrho} l \rceil - KL(n)$ times ;
    Update $t \leftarrow t + \lceil an^{\varrho} l \rceil$ and batch index $n \leftarrow n + 1$;
```

epoch length may lead to excessive exploitation based on an outdated estimate of the mean rewards. To address these issues, we modify the DSEE algorithm by using only the rewards from the current exploration epoch to estimate the mean rewards, and we increase the length of the exploitation epoch using a power law instead of an exponential function.

3.3 Analysis of the LM-DSEE Algorithm

Before we analyze the LM-DSEE algorithm, we introduce the following notation for the piece-wise stationary environment. Let

$$\Delta_k = \max\{\mu_t^* - \mu_t^k \mid t \in \{1, \ldots, T\}\},$$

$$\Delta_{\max} = \max\{\Delta_k \mid k \in \{1, \ldots, K\}\},$$

$$\text{and } \Delta_{\min} = \min\{\mu_t^* - \mu_t^k \mid t \in \{1, \ldots, T\}, k \in \{1, \ldots, K\}, \mu_t^* - \mu_t^k > 0\}.$$

Theorem 3.1 (Regret Upper Bound for LM-DSEE). *For piece-wise stationary environment with* *number of breakpoints* $\Upsilon_T \in O(T^\nu)$ *and* $\nu \in [0,1)$*, the regret for the LM-DSEE algorithm satisfies*

$$R_T^{\text{LM-DSEE}} \in O(T^{\frac{1+\nu}{2}} \ln T).$$

Proof. Let N be the index of the epoch containing the time-instant T, then the length of each epoch is at most $\lceil N^\varrho l \rceil$. Since breakpoints are located in at most Υ_T epochs, we can upper bound the regret from epochs containing breakpoints by

$$R_{\text{b}} \leq \Upsilon_T \lceil N^\varrho l \rceil \Delta_{\max}.$$

In the epochs containing no breakpoint, let R_{e} and R_{i} denote, respectively, the regret from exploration and exploitation epochs. Note that in such epochs, the mean reward from each arm does not change. In the n-th epoch with no breakpoint, we denote the maximum mean reward by $\mu^*_{\text{no-break}}(n)$ and the set of arms with maximum mean reward by $\mathcal{K}^*_{\text{no-break}}(n)$. Then, the regret in exploration epochs R_{e} satisfies,

$$R_{\text{e}} \leq \sum_{n=1}^{N} \sum_{k=1}^{K} \lceil \gamma \ln(n^\varrho l) \rceil \Delta_k \leq N \lceil \gamma \ln(N^\varrho l) \rceil \sum_{k=1}^{K} \Delta_k.$$

In exploitation epochs, regret is incurred if a sub-optimal arm is selected, and consequently, the regret in exploitation epochs R_{i} satisfies

$$R_{\text{i}} \leq \sum_{n=1}^{N} \sum_{k=1}^{K} \big[\lceil n^\varrho l \rceil - KL(n) \big] \mathbb{P}(\varphi_n^{\text{epch}} = k \notin \mathcal{K}^*_{\text{no-break}}(n)) \Delta_k, \tag{3.2}$$

where φ_n^{epch} is the arm selected in the n-th exploitation epoch.

It follows from the Chernoff-Hoeffding inequality [83, Theorem 1] that

$$\mathbb{P}\big(\bar{\mu}_k^{\text{epch}}(n) \geq \mu_k^{\text{epch}}(n) + \delta\big) = \mathbb{P}\big(\bar{\mu}_k^{\text{epch}}(n) \leq \mu_k^{\text{epch}}(n) - \delta\big) = \exp(-2\delta^2 L(n)),$$

where $\mu_k^{\text{epch}}(n)$ is the mean reward of arm k in the n-th epoch and $L(n)$ is the number of times an arm is selected in the n-th exploration epoch. Thus, we take $j^* \in \mathcal{K}^*_{\text{no-break}}(n)$ and get

$$\begin{aligned}
&\mathbb{P}\big(\varphi_n^{\text{epch}} = k \notin \mathcal{K}^*_{\text{no-break}}(n)\big) \\
\leq &\mathbb{P}\big(\bar{\mu}_k^{\text{epch}}(n) \geq \mu_k^{\text{epch}}(n) + \frac{\Delta_{\min}}{2}\big) + \mathbb{P}\big(\bar{\mu}_{j^*}^{\text{epch}}(n) \leq \mu^*_{\text{no-break}}(k) - \frac{\Delta_{\min}}{2}\big) \\
\leq &2 \exp\Big(- \frac{\Delta_{\min}^2}{2} \gamma \ln(n^\varrho l) \Big).
\end{aligned}$$

Since $\gamma \geq \frac{2}{\Delta_{\min}^2}$, $\mathbb{P}\big(\varphi_n^{\text{epch}} = k \notin \mathcal{K}^*_{\text{no-break}}(n)\big) \leq 2(n^{\varrho}l)^{-1}$. Substituting it into (3.2), we have $R_{\text{i}} \leq 2N\sum_{k=1}^{K}\Delta_k$ since $\lceil n^{\varrho}l\rceil - KL(n) < n^{\varrho}l$. Furthermore, it can be seen that

$$\frac{l}{1+\varrho}(N-1)^{1+\varrho} - N \leq T \leq \frac{l}{1+\varrho}(N+1)^{1+\varrho} + N,$$

and consequently $N \in O(T^{\frac{1}{1+\varrho}})$. Therefore, it follows that

$$R^{\text{LM-DSEE}}(T) = R_{\text{b}} + R_{\text{e}} + R_{\text{i}} \leq \Upsilon_T N^{\varrho} l \Delta_{\max} + N(\lceil \gamma \ln(N^{\varrho}l)\rceil + 2)\sum_{k=1}^{K}\Delta_k.$$

Thus, the regret $R^{\text{LM-DSEE}}(T) \in O(T^{\frac{1+\nu}{2}}\ln T)$, and this establishes the theorem. □

3.4 The SW-UCB# Algorithm

The SW-UCB# algorithm is an adaptation of the SW-UCB algorithm proposed and studied in [20]. At time t, SW-UCB# maintains an estimate of the mean reward $\bar{\mu}_k(t,\alpha)$ at each arm k, using only the rewards collected within a sliding-window of observations. Let the width of the sliding-window at time $t \in \{1,\ldots,T\}$ be $\tau(t,\alpha) = \min\{\lceil \lambda t^{\alpha}\rceil, t\}$, where parameters $\alpha \in (0,1]$, $\xi \in (1,2]$ and $\lambda \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$ are tuned based on environment characteristics. Let

$$n_k(t,\alpha) = \sum_{s=t-\tau(t,\alpha)+1}^{t} \mathbf{1}\{\varphi_s = k\}$$

be the number of times arm k has been selected in the time-window at time t, then we have

$$\bar{\mu}_k(t,\alpha) = \frac{1}{n_k(t,\alpha)} \sum_{s=t-\tau(t,\alpha)+1}^{t} X_s^{\varphi_s}\mathbf{1}\{\varphi_s = k\}.$$

Based on the above estimate, the SW-UCB# algorithm at each time selects the arm

$$\varphi_t = \operatorname*{arg\,max}_{k\in\{1,\ldots,K\}} \; \bar{\mu}_k(t-1,\alpha) + c_k(t-1,\alpha), \tag{3.3}$$

where $c_k(t,\alpha) = \sqrt{\xi \ln(t)/n_k(t,\alpha)}$. The details of the algorithm are presented in Algorithm 2.

In contrast to the SW-UCB algorithm [20], the SW-UCB# algorithm employs a time-varying width of the sliding-window. The tuning of the fixed window width in [20] requires a priori knowledge of the time horizon T which is no longer needed for the SW-UCB# algorithm.

Algorithm 2: The SW-UCB# Algorithm

```
Input   : $\nu \in [0,1)$, $\Delta_{\min} \in (0,1)$, $\lambda \in \mathbb{R}_{>0}$ & $T \in \mathbb{N}$;
Set     : $\alpha = \frac{1-\nu}{2}$
Output  : sequence of arm selection;
% Initialization:
while $t \leq T$ do
    if $t \in \{1, \dots, N\}$ then
        Pick arm $\varphi_t = t$;
    else
        Pick arm $\varphi_t$ defined in (3.3) ;
```

3.5 Analysis of the SW-UCB# Algorithm

We analyze the performance of the SW-UCB# algorithm (Algorithm 2) to get the following result.

Theorem 3.2 (Regret Upper Boudn for SW-UCB#)**.** *For the piece-wise stationary environment with number of breakpoints $\Upsilon_T = O(T^\nu)$ and $\nu \in [0,1)$, the regret for the SW-UCB# algorithm satisfies*

$$R_T^{\text{SW-UCB\#}} \in O(T^{\frac{1+\nu}{2}} \ln T).$$

Proof. We define set $\hat{\mathcal{T}}$ such that for all $t \in \hat{\mathcal{T}}$, t is either a breakpoint or there exists a break point in its sliding-window of observations $\{t - \tau(t-1, \alpha), \dots, t-1\}$. For $t \in \hat{\mathcal{T}}$, the statistical means are corrupted. Since the maximum sliding-window width is $\lceil \lambda (T-1)^\alpha \rceil$, it can be shown that

$$|\mathcal{T}| \leq \Upsilon_T \lceil \lambda (T-1)^\alpha \rceil.$$

Then, the regret can be upper bounded as follows.

$$R_T^{\text{SW-UCB\#}} \leq \Upsilon_T \lceil \lambda (T-1)^\alpha \rceil \Delta_{\max} + \sum_{k=1}^{K} \mathbb{E}[\tilde{N}_k(T)] \Delta_k, \tag{3.4}$$

where $\tilde{N}_k(T) := \sum_{t=1}^{T} \mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*,\ t \notin \hat{\mathcal{T}}\}$, and $\mathcal{K}_t^*$ is the set of arms with maximum mean reward at t. It can be seen that

$$\begin{aligned}\tilde{N}_k(T) \leq 1 + &\sum_{t=K+1}^{T} \mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*, n_k(t-1,\alpha) < A(t-1)\} \\ + &\sum_{t=K+1}^{T} \mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*,\, t \notin \hat{\mathcal{T}},\, n_k(t-1,\alpha) \geq A(t-1)\},\end{aligned} \tag{3.5}$$

where $A(t) = 4\xi \ln t/\Delta_{\min}^2$.

We first bound the second term on the right side of inequality (3.5). Let $G \in \mathbb{N}$ be such that

$$[\lambda(1-\alpha)(G-1)]^{\frac{1}{1-\alpha}} < T \leq [\lambda(1-\alpha)G]^{\frac{1}{1-\alpha}}. \tag{3.6}$$

Then, consider the following partition of time indices

$$\Big\{\big\{1 + \big\lfloor [\lambda(1-\alpha)(g-1)]^{\frac{1}{1-\alpha}} \big\rfloor, \ldots, \big\lfloor [\lambda(1-\alpha)g]^{\frac{1}{1-\alpha}} \big\rfloor\big\}\Big\}_{g \in \{1,\ldots,G\}}. \tag{3.7}$$

In the g-th epoch in the partition, either

$$\sum_{t \in g\text{-th epoch}} \mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*, n_k(t-1,\alpha) < A(t-1)\} = 0,$$

or there exist at least one time instant t that $\varphi_t = k \notin \mathcal{K}_t^*$ and $n_j(t-1,\alpha) < A(t-1)$. Let the last time instant satisfying these conditions in the g-th epoch be

$$t_k(g) = \max\big\{t \in g\text{-th epoch} \mid \varphi_t = k \notin \mathcal{K}_t^* \text{ and } n_k(t-1,\alpha) < A(t)\big\}.$$

We will now show that there exists at most one time index in the g-th epoch until $t_k(g) - 1$ that is not covered by the time-window at $t_k(g)$. Towards this end, consider the increasing convex function $f(x) = x^{\frac{1}{1-\alpha}}$ with $\alpha \in (0,1)$. It follows that $f(x_2) - f(x_1) \leq f'(x_2)(x_2 - x_1)$ if $x_2 \geq x_1$. Let $\tilde{t}$ be a time index in the g-th epoch, and set $x_1 = g-1$ and $x_2 = \frac{\tilde{t}^{1-\alpha}}{\lambda(1-\alpha)}$. Then, substituting x_1 and x_2 in the above inequality and simplifying, we get

$$\tilde{t} - (\lambda(1-\alpha)(g-1))^{\frac{1}{1-\alpha}} \leq \lambda\tilde{t}^{\alpha}\Big(\frac{\tilde{t}^{1-\alpha}}{\lambda(1-\alpha)} - g + 1\Big). \tag{3.8}$$

Since by definition of the g-th epoch, $\frac{\tilde{t}^{1-\alpha}}{\lambda(1-\alpha)} \leq g$, we have

$$\tilde{t} - \lfloor(\lambda(1-\alpha)(g-1))^{\frac{1}{1-\alpha}}\rfloor \leq \min\{\tilde{t}+1, \lambda\lceil\tilde{t}^{\alpha}\rceil + 1\} = \tau(\tilde{t},\alpha) + 1. \tag{3.9}$$

Setting $\tilde{t} = t_k(g) - 1$ in (3.9), we obtain

$$t_k(g) - \tau\big(t_k(g) - 1, \alpha\big) \leq 2 + \lfloor \lambda(1-\alpha)(g-1)^{\frac{1}{1-\alpha}} \rfloor,$$

i.e., the first time instant in the sliding-window at $t_j(g)$ is located at or to the left of the second time instant of the g-th epoch in the partition (3.7). Therefore, it follows that

$$\begin{aligned}&\sum_{t=1+\lfloor \lambda(1-\alpha)(g-1)^{\frac{1}{1-\alpha}} \rfloor}^{\lfloor \lambda(1-\alpha) g^{\frac{1}{1-\alpha}} \rfloor} \mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*, n_k(t-1,\alpha) < A(t-1)\} \\ &\leq n_k(t-1,\alpha) + 2 \leq A(t_k(g)-1) + 2.\end{aligned}$$

Now we have

$$\begin{aligned}&\sum_{t=K+1}^{T} \mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*, n_k(t-1,\alpha) < A(t-1)\} \\ \leq& 2G + \sum_{g=1}^{G} A(t_k(g)-1) \leq G\Big(2 + \frac{4\xi \ln T}{\Delta_{\min}^2}\Big).\end{aligned} \tag{3.10}$$

Next, we upper-bound the expectation of the last term on the right-hand side of inequality (3.5). Taking $j_t^* \in \mathcal{K}_t^*$, it can be shown that

$$\begin{aligned}&\mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*, t \notin \hat{\mathcal{T}}, n_k(t-1,\tau) \geq A(t-1)\} \\ \leq& \sum_{s^*=1}^{\lceil \lambda(t-1)^{\alpha} \rceil} \sum_{s=A(t-1)}^{\lceil \lambda(t-1)^{\alpha} \rceil} \mathbf{1}\{n_k(t-1,\alpha) = s,\ n_{j_t^*}(t-1,\alpha) = s^*\} \\ &\times \mathbf{1}\{\bar{\mu}_k(t-1,\alpha) + c_k(t-1,\alpha) > \bar{\mu}_{j_t^*}(t-1,\alpha) + c_{j_t^*}(t-1,\alpha),\ t \notin \hat{\mathcal{T}}\}.\end{aligned} \tag{3.11}$$

When $t \notin \hat{\mathcal{T}}$, for each arm $k \in \{1, \dots, K\}$, $\mu_k(s)$ is a constant for all $s \in \{t - \tau(t-1,\alpha), \dots, t\}$. Note that if $\bar{\mu}_k(t-1,\alpha) + c_k(t-1,\alpha) > \bar{\mu}_{j_t^*}(t-1,\alpha) + c_{j_t^*}(t-1,\alpha)$ is true, at least one of the following inequalities holds.

$$\bar{\mu}_k(t-1,\alpha) \geq \mu_t^k + c_k(t-1,\alpha), \tag{3.12}$$

$$\bar{\mu}_{j_t^*}(t-1,\alpha) \leq \mu_t^{j_t^*} - c_{j_t^*}(t-1,\alpha), \tag{3.13}$$

$$\mu_t^* - \mu_t^k < 2c_k(t-1,\alpha). \tag{3.14}$$

Since $n_k(t-1,\alpha) \geq A(t-1)$, (3.14) does not hold. Applying Chernoff-Hoeffding inequality [83, Theorem 1] to bound the probability of events (3.12) and (3.13), we obtain

$$\mathbb{P}(\bar{\mu}_k(t-1,\alpha) \geq \mu_k(t) + c_k(t-1,\alpha)) \leq (t-1)^{-2\xi},$$
$$\mathbb{P}(\bar{\mu}_{j_t^*}(t-1,\alpha) \leq \mu_{j_t^*}(t) - c_{j_t^*}(t-1,\alpha)) \leq (t-1)^{-2\xi},$$

where $\xi = 1+\alpha$. Applying both probability inequalities in conjuncture with (3.11), we get

$$\begin{aligned}&\mathbb{E}\Big[\sum_{t=K+1}^{T} \mathbf{1}\{\varphi_t = k \notin \mathcal{K}_t^*,\, t \notin \hat{\mathcal{T}},\, n_k(t-1,\alpha) \geq A(t-1)\}\Big]\\ &\leq \sum_{t=K+1}^{T} 2(t-1)^{-2\xi}[\lambda(t-1)^{\alpha}+1]^2\\ &\leq \sum_{t=1}^{\infty} 2(\lambda+1)^2 t^{-2} = \frac{(\lambda+1)^2\pi^2}{3}. \qquad (3.15)\end{aligned}$$

Therefore, it follows from (3.4), (3.5), (3.10), and (3.15) that

$$R_T^{\text{SW-UCB\#}} \leq \Upsilon_T \lceil \lambda(T-1)^{\alpha} \rceil \Delta_{\max} + \sum_{k=1}^{K}\left(G\Big(2+\frac{4\xi \ln T}{\Delta_{\min}^2}\Big) + 1 + \frac{(\lambda+1)^2\pi^2}{3}\right)\Delta_k.$$

From (3.6), we have $G = O(T^{1-\alpha})$, and this yields $R_T^{\text{SW-UCB\#}} \in O(T^{\frac{1+\nu}{2}} \ln T)$. □

3.6 Numerical Illustration

In this section, we present simulation results for the SW-UCB# and LM-DSEE algorithms. For each simulation, we consider a 10-armed bandit in which the reward at each arm is generated using Beta distribution. The breakpoints are introduced at time instants where the next element of the sequence $\{\lfloor t^{\nu} \rfloor\}_{t\in\{1,\ldots,T\}}$ is different from the current element. At each breakpoint, the mean rewards at each arm were randomly selected from the set $\{0.05, 0.12, 0.19, 0.26, 0.33, 0.39, 0.46, 0.53, 0.6, 0.9\}$. We select the parameters (a, b) equal to $(1, 0.25)$ for LM-DSEE in Algorithm 1. For SW-UCB# in Algorithm 2, we select $\lambda = 12.3$.

The parameters ν that describe characteristics of nonstationarity are varied to evaluate the performance of algorithms. Figure. 3.1 shows that both SW-UCB# and LM-DSEE are effective in the piece-wise stationary environment.

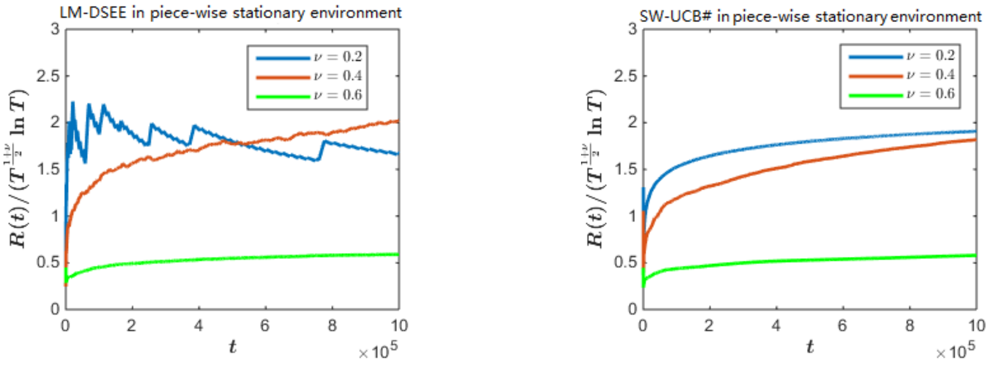


Figure 3.1: Comparison of LM-DSEE and SW-UCB#.

It can be seen in Figs. 3.1 that for both algorithms, as expected, the ratio of the empirical regret to the order of the regret established in Sections 3.3 and 3.5 is upper bounded by a constant. The regret for the SW-UCB# is relatively smoother than the regret for the LM-DSEE algorithm. The saw-tooth behavior of the regret for LM-DSEE is attributed to the fixed exploration-exploitation structure, wherein the regret is certainly incurred during the exploration epochs.

3.7 Summary

We studied the stochastic MAB problem in the piece-wise stationary environment and designed two novel algorithms, the LM-DSEE and the SW-UCB# for these problems. We analyzed these algorithms to show that these algorithms incur sublinear regret, i.e., the time average of the regret asymptotically converges to zero. The theoretical results are verified with numerical illustrations.

While both the algorithms incur the same order of regret, compared with LM-DSEE, SW-UCB# has a better leading constant. This illustrates the cost of constraining the algorithm to have a deterministic structure. On the other hand, this deterministic structure can be very useful, for example, in the context of planning trajectories for a mobile robot performing surveillance or search using a MAB framework. Though both algorithms can balance the explore-exploit tradeoff, they are reactive in the sense that they select only one arm at a time, i.e., they only provide information about the next location to be visited by the robot. Certain motion constraints on the robots such as non-holonomicity may make such movements energetically demanding. Therefore, the deterministic and predictable structure of LM-DSEE can be leveraged to design a tour for the robot which can be efficiently traversed even under motion constraints.

There are several possible extensions of this work. In the next chapter, we'll study the multiple decision-maker version of the problem in this chapter. Besides, extensions of the methodology developed in this paper to other classes on MAB problems such as the Markovian MAB problem [84] and the restless bandits [85] are also of interest.

3.8 Bibliographic Remarks

In a non-stationary environment, achieving logarithmic expected cumulative regret may not be feasible and the focus is on the design of algorithms that achieve sub-linear regret. Indeed, the lower bound for the piece-wise stationary stochastic bandit has been shown to be $\Omega(\sqrt{K\Upsilon_T T})$ in [20]. Thus, both of the proposed algorithms in this chapter are near-optimal. The approaches to handle nonstationary environments can be classified into active approach and passive approach. The former actively detects the breakpoints to and accordingly removes the old sampling results, while the latter follows a predetermined rule and disregards the information about breakpoints.

One of the passive approaches is proposed by Kocsis and Szepesvári [86] which uses a discounting factor to compute the UCB index. In subsequent work, Garivier and Moulines [20] provide a formal analysis of Discounted UCB (D-UCB) and propose SW-UCB, which is also a passive approach. They pointed out that if the number of change points Υ_T is available, both algorithms can be tuned to achieve a regret close to the $\Omega(\sqrt{K\Upsilon_T T})$ regret lower bound.

The active approach handles the change of reward distributions in an adaptive manner. Harland et al. [87] actively detect the change point with the Page-Hinkley test and design two restarting strategies to prevent false alarm, namely γ-Restart and Meta-Bandit. The γ-Restart triggers discounting the sampling history when a breakpoint is detected, while the meta-Bandit models preserving or discarding old information as a new 2-armed bandit. Other change point detection techniques such as cumulative sum (CUMSUM) and Generalized Likelihood Ratio (GLR) tests are used in subsequent work to design CUMSUM-UCB [88], GLR-klUCB [89], and M-UCB [90]. Change point detection has also been implemented together with non-UCB methods to design EXP3.R [91] and Change-Point Thompson sampling [92]. These policies either need the knowledge of Υ_T to tune the parameter or have regret upper bounds with a strong dependence on N_T. Very recently, two parameter-free policies AdSwitch [93] and Ada-ILTCB$^+$ [94] for contextual MAB are proved to have $\tilde{O}(\sqrt{KN_TT})$ regret.

CHAPTER 4

MULTI-PLAYER PIECEWISE STATIONARY STOCHASTIC BANDITS

In a variety of applications including robotic swarming, opportunistic spectrum access, and the Internet of Things, achieving coordinated behavior of multiple decision-makers in unknown, uncertain, and non-stationary environments without any explicit communication among them is of immense interest. This multi-player decision-making in the face of uncertainty is embodied by the multi-player MAB problem, in which several decision-makers simultaneously play the bandit game in a decentralized fashion. For such a problem, we follow a common routine to assume a collision model: the reward from an arm is eliminated or shared when it is selected by multiple agents. The work in this chapter is slightly modified from our published paper on multi-player piecewise stationary stochastic bandits, and it is reproduced here with the permission of the copyright holder[1].

To formally formulate the problem, we consider a multi-player MAB problem with K arms and the total number of players is $M \in \{1, \dots, K\}$. Similarly as the single-player case in last chapter, at each time t, there is a random reward $X_t^k \in [0, 1]$ associated with each arm $k \in \{1, \dots, K\}$, and every agent $j \in \{1, \dots, M\}$ picks a particular arm $\varphi_t(j) \in \{1, \dots, K\}$ and observes $X_t^{\varphi_t(j)}$. We assume no communication between agents, so that $\varphi_t(j)$ is selected based only on agent j's own observation and decision-making history $\left\{X_s^{\varphi_s(j)}, \varphi_s(j)\right\}_{s=1}^{t-1}$.

With collision model $\mathcal{M}$ that eliminate rewards, agent j receives the reward $X_t^{\varphi_t(j)}$ from arm $\varphi_t(j)$ if it is the only player to select arm $\varphi_t(j)$ at time t. Then, the *group reward* till time T is

$$S_T = \sum_{t=1}^{T} \sum_{k=1}^{K} X_t^k O_t^k,$$

where $O_t^k = 1$ if arm k is selected by only one player at time t and is zero otherwise. If the collision model allows the reward to be shared, $O_t^k = 1$ if arm k is selected by a player. Since the algorithm design and analysis are similar in both cases, the discussion will only be made based on the collision model $\mathcal{M}$ in this chapter.

[1] ©2018 IEEE. Reprinted with permission from [95].

We assume the minimum difference in mean rewards between any pair of arms at any time is lower bounded by $\Delta_{\min} > 0$. Let σ_t be a permutation of $\{1, \ldots, K\}$ at time t such that the mean rewards satisfy

$$\mu_t^{\sigma_t(1)} > \ldots > \mu_t^{\sigma_t(K)}.$$

Then, the *group regret* for a policy ρ till time T is defined by

$$R_T^\rho(\mathcal{M}) = \sum_{t=1}^{T} \sum_{k=1}^{M} \mu_t^{\sigma_t(k)} - \mathbb{E}^\rho\left[S_T\right] = \sum_{t=1}^{T} \sum_{k=1}^{M} \mu_t^{\sigma_t(k)} - \mathbb{E}^\rho\left[\sum_{t=1}^{T} \sum_{k=1}^{K} \mu_t^k O_t^k\right],$$

where the second expectation is computed over different realizations of O_t^k under policy ρ. Our main purpose here is to design a multi-player policy ρ that minimizes $R_T^\rho(\mathcal{M})$. Like the last chapter, we study the above MAB problem in a piecewise stationary environment with the number of breakpoints until time T to be $\Upsilon_T \in O(T^\nu)$, where $\nu \in [0, 1)$ is known a priori.

4.1 The RR-SW-UCB# Algorithm

The Round Robin SW-UCB# (RR-SW-UCB#) algorithm is designed based upon SW-UCB# presented in the last chapter. In the RR-SW-UCB# algorithm, at each time t, every agent j maintains an estimate of the mean reward $\bar{\mu}_k^j(t, \alpha)$ at each arm k, using only the rewards collected within a sliding-window of width $\tau(t, \alpha) = \min\{\lceil \lambda t^\alpha \rceil, t\}$, where parameter $\alpha \in (0, 1]$. The number of times arm k has been selected within the time-window at time t is

$$n_k^j(t, \alpha) = \sum_{s=t-\tau(t,\alpha)+1}^{t} \mathbf{1}\{\varphi_s(j) = k\}.$$

Then, $\bar{\mu}_k^j(t, \alpha)$ can be computed by

$$\bar{\mu}_k^j(t, \alpha) = \frac{1}{n_k(t, \alpha)} \sum_{s=t-\tau(t,\alpha)+1}^{t} X_s^{\varphi_s(j)} \mathbf{1}\{\varphi_s(j) = k\}.$$

Using its own observations, each agent j computes upper confidence bounds on the mean rewards

$$\bar{\mu}_k^j(t-1, \alpha) + c_k^j(t-1, \alpha), \quad \forall k \in \{1, \ldots, K\},$$

where $c_k^j(t-1, \alpha) = \sqrt{(1+\alpha) \ln t / n_k^j(t-1, \alpha)}$. For initial K iterations, i.e., $t \in \{1, \ldots, K\}$, the player j selects each arm once. Then, at time instants $\{K + \eta M + 1\}_{\eta \in \mathbb{Z}_{\geq 0}}$, it computes the set Ω_j

Algorithm 3: The RR-SW-UCB# Algorithm

```
Input   : $\nu \in [0,1)$, $\Delta_{\min} \in (0,1)$, $\lambda \in \mathbb{R}_{>0}$ , $T \in \mathbb{N}$ and player number $j$;
Set     : $\alpha = \frac{1-\nu}{2}$;
Output  : sequence of arm selections for each player $j$;
% Initialization:
Set $\Omega_j \leftarrow \emptyset$, ordered set $\mathcal{G}_j \leftarrow ()$, and $t \leftarrow 1$;
while $t \leq T$ do
    % round-robin selection of each arm starting at arm j
    if $t \in \{1, \ldots, K\}$ then
        Pick arm $\varphi_t(j) = \mathrm{mod}(t + j - 2, K) + 1$;
    else
        Compute $\Omega_j$ containing $M$ arms with $M$ largest values in
            $\left\{\bar{\mu}_k^j(t-1,\alpha) + c_k^j(t-1,\alpha) \mid k \in \{1,\ldots,K\}\right\}$;
        Ascending sort the arm indices in $\Omega_j$, $\mathcal{G}_j \leftarrow \mathrm{sort}_\uparrow(\Omega_j)$;
        % round-robin selection of arms in $\mathcal{G}_j$ starting at $\mathcal{G}_j(j)$
        for round $\in \{1, \ldots, M\}$ do
            Pick arm $\varphi_t(j) = \mathcal{G}_j(\mathrm{mod}(t - K + j - 2, M) + 1)$;
            $t \leftarrow t + 1$;
```

containing M arms with M largest values in the set

$$\left\{\bar{\mu}_k^j(t-1,\alpha) + c_k^j(t-1,\alpha) \mid k \in \{1,\ldots,K\}\right\}.$$

Let $\mathcal{G}_j$ be the ordered set that contains arms in Ω_j sorted in ascending value of their indices (not using the upper confidence bounds), and let $\mathcal{G}_j(i)$ denote the i-th element in $\mathcal{G}_j$. The player j selects arms in $\mathcal{G}_j$ in a round-robin fashion starting with the arm $\mathcal{G}_t^j(j)$. It will be shown in the following section that the estimated set of M best arms, denoted by Ω_j, will be the same for each player with high probability. Details of RR-SW-UCB# are shown in Algorithm 3. The free parameter λ in the algorithm can be used to refine the finite-time performance of the algorithm.

4.2 Analysis of the RR-SW-UCB# Algorithm

Before the analysis, we introduce the following notation. Let $\Omega_*^M(t)$ denote the set of M arms with the M largest mean rewards at time t. Then, the total number of times $\Omega_j(t) \neq \Omega_*^M(t)$ until time T

can be defined as

$$\mathcal{N}_j(T) := \sum_{t=1}^{T} \mathbf{1}\{\Omega_j(t) \neq \Omega_*^M(t)\}.$$

We now upper bound $\mathcal{N}_j(T)$ in the following lemma.

Lemma 4.1. *For the RR-SW-UCB# algorithm and the multi-player MAB problem with K arms and M players in the piecewise stationary environment with the number of break points* $\Upsilon_T = O(T^\nu)$, $\nu \in [0, 1)$, *the total number of times* $\Omega_j(t) \neq \Omega_*^M(t)$ *until time T for any player k satisfies*

$$\begin{aligned}\mathcal{N}_j(T) \leq & (K-M)\Big[\Big(\frac{T^{1-\alpha}}{\lambda(1-\alpha)}+1\Big)\Big(1+\frac{4M(1+\alpha)\ln T}{\Delta_{\min}^2}\Big)+\frac{\pi^2}{3}\Big(\frac{\lambda+M+1}{M}\Big)^2\Big] \\ & + \Upsilon_T\big(\lceil \lambda(T-1)^\alpha \rceil + M - 1\big) + K.\end{aligned}$$

Proof. We begin by separately analyzing windows with and without breakpoints. For the ease of notation, in the following, superscript j is omitted in $\bar{\mu}_k^j(t,\alpha)$, $n_k^j(t,\alpha)$ and $c_k^j(t,\alpha)$.

Step 1: Let set $\hat{\mathcal{T}}$ such that for all $t \in \hat{\mathcal{T}}$, t is either a breakpoint or there exists a break point in its sliding-window of observations $\{t - \tau(t-1,\alpha), \ldots, t-1\}$. For $t \in \hat{\mathcal{T}}$, the statistical means are biased. It follows that

$$|\hat{\mathcal{T}}| \leq \Upsilon_T \lceil \lambda(T-1)^\alpha \rceil.$$

Consequently, $\mathcal{N}_j(T)$ can be upper-bounded as

$$\mathcal{N}_j(T) \leq \Upsilon_T\Big(\lceil \lambda(T-1)^\alpha \rceil + M - 1\Big) + \tilde{\mathcal{N}}_j(T), \tag{4.1}$$

where $\tilde{\mathcal{N}}_j(T) := \sum_{t=1}^{T} \mathbf{1}\{\Omega_j(t) \neq \Omega_*^M(t), t \notin \hat{\mathcal{T}}\}$. The term $M-1$ in (4.1) is due to the fact that Ω_j is computed every M steps. In the following steps, we will bound $\tilde{\mathcal{N}}_j(T)$.

Step 2: If $\Omega_j(t) \neq \Omega_*^M(t)$, there exists at least one arm i such that $i \in \Omega_j(t)$ and $i \notin \Omega_*^M(t)$. Then, it follows that

$$\begin{aligned}\tilde{\mathcal{N}}_j(T) \leq & K + \sum_{t=K+1}^{T}\sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t), i \notin \Omega_*^M(t), t \notin \hat{\mathcal{T}}, n_i(t-1,\alpha) < l(t,\alpha)\} \\ & + \sum_{t=K+1}^{T}\sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t),\ i \notin \Omega_*^M(t), t \notin \hat{\mathcal{T}}, n_i(t-1,\alpha) \geq l(t,\alpha)\},\end{aligned} \tag{4.2}$$

where we choose $l(t,\alpha) = 4(1+\alpha)\ln t/\Delta_{\min}^2$.

We begin with bounding the second term on the right-hand side of inequality (4.2). First, we partition time instants into G epochs. Let $G \in \mathbb{N}$ be such that

$$[\lambda(1-\alpha)(G-1)]^{\frac{1}{1-\alpha}} < T \le [\lambda(1-\alpha)G]^{\frac{1}{1-\alpha}}. \tag{4.3}$$

Then, we have the following epochs

$$\{1+\phi(g-1), \dots, \phi(g)\}_{g\in\{1,\dots,G\}}, \tag{4.4}$$

where $\phi(g) = \left\lfloor [\lambda(1-\alpha)g]^{\frac{1}{1-\alpha}} \right\rfloor$. Let $\tilde{t}$ be any time instant other than the first instant in the g-th epoch. We will now show that all but one of the time instants in the g-th epoch until $\tilde{t}$ must be contained in the time-window at $\tilde{t}$. Towards this end, consider the increasing convex function $f(x) = x^{\frac{1}{1-\alpha}}$ with $\alpha \in (0,1)$. It follows that $f(x_2) - f(x_1) \le f'(x_2)(x_2 - x_1)$ if $x_2 \ge x_1$. Then, substituting $x_1 = g-1$ and $x_2 = \frac{\tilde{t}^{1-\alpha}}{\lambda(1-\alpha)}$ in the above inequality and simplifying, we get

$$\tilde{t} - (\lambda(1-\alpha)(g-1))^{\frac{1}{1-\alpha}} \le \lambda \tilde{t}^{\alpha}\Big(\frac{\tilde{t}^{1-\alpha}}{\lambda(1-\alpha)} - g + 1\Big).$$

Since by definition of the g-th epoch, $\frac{\tilde{t}^{1-\alpha}}{\lambda(1-\alpha)} \le g$, we have

$$\tilde{t} - \lfloor(\lambda(1-\alpha)(g-1))^{\frac{1}{1-\alpha}}\rfloor \le \min\{\tilde{t}+1, \lambda\lceil\tilde{t}^{\alpha}\rceil + 1\} = \tau(\tilde{t},\alpha) + 1.$$

The only time instant in the g-th epoch that is possibly not contained in the time window at $\tilde{t}$ is $1+\phi(g-1)$. Then for any arm $i \in \{1,\dots,K\}$,

$$\sum_{2+\phi(g-1)}^{\tilde{t}} \mathbf{1}\{i \in \Omega_j(t)\} \le M n_i(\tilde{t},\alpha). \tag{4.5}$$

Furthermore, in the g-th epoch in the partition, either

$$\sum_{t\in g\text{-th epoch}} \sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t),\, i \notin \Omega_*^M(t),\, t \notin \hat{\mathcal{T}},\, n_i(t-1,\alpha) < l(t,\alpha)\} = 0,$$

or there exist at least one time-instant t in the g-th epoch such that

$$\sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t),\, i \notin \Omega_*^M(t),\, t \notin \hat{\mathcal{T}},\, n_i(t-1,\alpha) < l(t,\alpha)\} > 0.$$

Let the last time instant satisfying this condition in the g-th epoch be

$$t(g) = \max\left\{t \in g\text{-th epoch} \mid \sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t),\, i \notin \Omega_*^M(t),\, t \notin \hat{\mathcal{T}},\, n_i(t-1,\alpha) < l(t,\alpha)\} > 0\right\}.$$

Note that $t(g) \notin \hat{\mathcal{T}}$ indicates, for each $i \in \{1,\dots,K\}$, $\mu_i(s)$ is a constant for all $s \in \{t(g) - \tau(t(g) - 1,\alpha),\dots,t(g)\}$. Then, it follows from (4.5) that

$$\begin{aligned}
&\sum_{t\in g\text{-th epoch}} \sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t),\, i \notin \Omega_*^M(t), t \notin \hat{\mathcal{T}}, n_i(t-1,\alpha) < l(t,\alpha)\}\\
&\le K - M + \sum_{t=\phi(g-1)+2}^{t(g)} \sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t), i \notin \Omega_*^M(t), t \notin \hat{\mathcal{T}}, n_i(t-1,\alpha) < l(t,\alpha)\}\\
&\le K - M + \sum_{i\notin\Omega_*^M(t(g))} Ml(t_i(g),\alpha)\\
&\le (K-M)\Big(1 + \frac{4M(1+\alpha)\ln T}{\Delta_{\min}^2}\Big), \qquad (4.6)
\end{aligned}$$

where $t_i(g) = \max\{t \in g\text{-th epoch} \mid i \in \Omega_j(t),\, i \notin \Omega_*^M(t),\, t \notin \hat{\mathcal{T}},\, n_i(t-1,\alpha) < l(t,\alpha)\}$ and $t_i(g) \le t(g)$ for all $i \in \{1,\dots,K\}$. Therefore, from (4.4) and (4.6), we have

$$\begin{aligned}
&\sum_{t=N+1}^{T} \sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t), i \notin \Omega_*(t), n_i(t-1,\alpha) < l(t,\alpha)\}\\
\le& G(K-M)\Big(1 + \frac{4M(1+\alpha)\ln T}{\Delta_{\min}^2}\Big). \qquad (4.7)
\end{aligned}$$

Step 3: In this step, we bound the expectation of the last term in (4.2). It can be shown that

$$\begin{aligned}
&\sum_{i=1}^{K} \mathbf{1}\{i \in \Omega_j(t), i \notin \Omega_*^M(t), t \notin \hat{\mathcal{T}}, n_i(t-1,\alpha) \ge l(t,\alpha)\}\\
\le& \sum_{i\notin\Omega_*^M(t)} \sum_{\zeta\in\Omega_*^M(t)} \sum_{s_\zeta=1}^{h(t)} \sum_{s_i=l(t,\alpha)}^{h(t)} \mathbf{1}\{n_\zeta(t-1,\alpha) = s_\zeta,\, n_i(t-1,\alpha) = s_i, t \notin \hat{\mathcal{T}}\} \qquad (4.8)\\
&\times \mathbf{1}\{\bar{\mu}_\zeta(t-1,\alpha) + c_\zeta(t-1,\alpha) \le \bar{\mu}_i(t-1,\alpha) + c_i(t-1,\alpha), n_i(t-1,\alpha) \ge l(t,\alpha)\},
\end{aligned}$$

where $h(t) := \lceil \lceil \lambda(t-1)^\alpha \rceil / M \rceil$ is the maximum number of times an arm can be selected within the time window at $t-1$. Note that $\bar{\mu}_\zeta(t-1,\alpha) + c_\zeta(t-1,\alpha) \le \bar{\mu}_i(t-1,\alpha) + c_i(t-1,\alpha)$ means at least one of the following holds.

$$\bar{\mu}_i(t-1,\alpha) \geq \mu_i(t) + c_i(t-1,\alpha), \tag{4.9}$$

$$\bar{\mu}_\zeta(t-1,\alpha) \leq \mu_\zeta(t) - c_\zeta(t-1,\alpha), \tag{4.10}$$

$$\mu_\zeta(t) - \mu_i(t) < 2c_i(t-1,\alpha). \tag{4.11}$$

Since $n_i(t-1,\alpha) \geq l(t,\alpha)$, (4.11) does not hold. Applying Chernoff-Hoeffding inequality [83, Theorem 1] to bound the probability of events (4.9) and (4.10), we obtain

$$\mathbb{P}(\bar{\mu}_i(t-1,\alpha) \geq \mu_i(t) + c_i(t-1,\alpha)) \leq t^{-2(1+\alpha)}, \tag{4.12}$$

$$\mathbb{P}(\bar{\mu}_\zeta(t-1,\alpha) \leq \mu_\zeta(t) - c_\zeta(t-1,\alpha)) \leq t^{-2(1+\alpha)}. \tag{4.13}$$

Since Ω_j is only computed at time instants $\{K+\eta M+1\}_{\eta\in\mathbb{Z}_{\geq 0}}$, it follows from (4.8), (4.12) and (4.13) that

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=K+1}^{T}\sum_{i=1}^{K}\mathbf{1}\{i\in\Omega_j(t), i\notin\Omega_*^M(t), t\notin\hat{\mathcal{T}}, n_i(t-1,\alpha)\geq l(t,\alpha)\}\Big]\\
&\leq (K-M)M\sum_{s_\zeta=1}^{h(f(\eta))}\sum_{s_i=l(t,\alpha)}^{h(f(\eta))}\sum_{\eta=0}^{\lceil\frac{T-N}{M}\rceil} 2Mf(\eta)^{-2(1+\alpha)}\\
&\leq (K-M)M^2\sum_{\eta=0}^{\lceil\frac{T-N}{M}\rceil} 2f(\eta)^{-2(1+\alpha)}h(f(\eta))^2\\
&\leq (K-M)\Big(\frac{\lambda+M+1}{M}\Big)^2\sum_{\eta=1}^{\infty}2\eta^{-2}\\
&= \frac{\pi^2}{3}(K-M)\Big(\frac{\lambda+M+1}{M}\Big)^2,
\end{aligned} \tag{4.14}$$

where $f(\eta) := K+\eta M+1$. Therefore, it follows from (4.1), (4.2), (4.7), and (4.14) that

$$\mathcal{N}_j(T) \leq K + (K-M)\Big[G\Big(1+\frac{4M(1+\alpha)\ln T}{\Delta_{\min}^2}\Big)+\frac{\pi^2}{3}\Big(\frac{\lambda+M+1}{M}\Big)^2\Big] + \Upsilon_T\big(\lceil\lambda(T-1)^\alpha\rceil + M - 1\big).$$

From (4.3), we have $G \leq T^{1-\alpha}/(\lambda-\lambda\alpha)+1$, and this yields the desired result. □

Based on Lemma 4.1, we now establish the order of expected cumulative group regret of RR-SW-UCB# in the abruptly changing environment.

Theorem 4.2. *For the RR-SW-UCB# algorithm and the multi-player MAB problem with N arms and M players in the piecewise stationary environment with the number of break points $\Upsilon_T = O(T^\nu)$, $\nu \in [0, 1)$, under collision model $\mathcal{M}$, the expected cumulative group regret satisfies*

$$R_T^{\text{RR-SW-UCB\#}}(\mathcal{M}) \in O\big(T^{\frac{1+\nu}{2}} \ln T\big).$$

Proof. If all player identify $\Omega_*^M(t)$ correctly at time t, no expected regret is accrued. It follows from Lemma 4.1 that $\mathcal{N}_j(T) \in O(T^{\frac{1+\nu}{2}} \ln T)$ for all $j \in \{1, \dots, M\}$. The total number of times that any player misidentifies $\Omega_*^M(t)$ until time T can be upper bounded by $\sum_{j=1}^M \mathcal{N}_j(T)$. Thus, we conclude the proof. □

4.3 The SW-DLP Algorithm

Distributed Learning with Prioritization (DLP) [23] is designed for the multi-player stochastic MAB problem in a stationary environment. The idea of DLP is to assign player j to collect rewards from j-th best arm for most of circumstances. In a piecewise stationary environment, we extend the DLP algorithm to design SW-DLP using a sliding observation window. The upper confidence bounds on the mean rewards in SW-DLP are computed the same as SW-UCB#. SW-DLP employes an identical allocation rule as DLP, i.e., at each time instant t, player j computes a set $A_j(t)$ containing j arms with j largest values in the set

$$\big\{\bar{\mu}_k^j(t-1, \alpha) + c_k^j(t-1, \alpha) \mid k \in \{1, \dots, K\}\big\},$$

and selects arm

$$\varphi_t(j) = \operatorname*{arg\,min}_{k \in A_j(t)} \{\bar{\mu}_k^j(t-1, \alpha) - c_k^j(t-1, \alpha)\}.$$

Details of the SW-DLP is shown in Algorithm 4. The parameters in the SW-DLP algorithm are the same as in the RR-SW-UCB# algorithm. In the following, we will refer to $\bar{\mu}_k^j(t-1, \alpha) - c_k^j(t-1, \alpha)$ as the lower confidence bound on the estimate reward from arm k.

4.4 Analysis of the SW-DLP Algorithm

We analyze the performance of the SW-DLP algorithm (Algorithm 4) to get the following result.

Algorithm 4: The SW-DLP Algorithm

```
Input, output, and parameters are the same as RR-SW-UCB#
while $t \leq T$ do
    if $t \in \{1, \dots, K\}$ then
        Pick arm $\varphi_t(j) = \mathrm{mod}(t + j - 2, N) + 1$;
    else
        Compute $A_j(t)$ containing $j$ arms with $j$ largest values in
            $\{\bar{\mu}_k^j(t-1, \alpha) + c_k^j(t-1, \alpha) \mid k \in \{1, \dots, K\}\}$;
        Pick arm
            $\varphi_t(j) = \arg\min_{k \in A_j} \{\bar{\mu}_k^j(t-1, \alpha) - c_k^j(t-1, \alpha)\}$;
```

Theorem 4.3. *For the SW-DLP algorithm and the multi-player MAB problem with K arms and M players in the piecewise stationary environment with the number of break points $\Upsilon_T = O(T^\nu)$, $\nu \in [0, 1)$, under collision model $\mathcal{M}$, the expected cumulative group regret satisfies*

$$R_T^{\text{SW-DLP}}(\mathcal{M}) \in O\big(T^{\frac{1+\nu}{2}} \ln T\big).$$

Proof. The proof is similar to the proof of Theorem 4.2 and we only present a sketch. Let $\theta_t(j)$ be the j-th best arm at time t. The total number of time instants that $\theta_t(j)$ is not selected by player j with SW-DLP satisfies

$$\hat{\mathcal{N}}_j = \sum_{t=1}^{T} \mathbf{1}\{\varphi_t(j) \neq \theta_t(j)\} \leq \Upsilon_T \lceil \lambda (T-1)^\alpha \rceil + \sum_{t=1}^{T} \mathbf{1}\{\varphi_t(j) \neq \theta_t(j),\, t \notin \hat{\mathcal{T}}\}. \tag{4.15}$$

We partition the time horizon as in (4.4). Then, similarly to (4.5) in the proof of Lemma 4.1, it can be shown that

$$\sum_{t=2+\phi(g-1)}^{\tilde{t}} \mathbf{1}\{\varphi_t(j) = i\} \leq n_i^j(\tilde{t}, \alpha), \tag{4.16}$$

for any arm $i \in \{1, \dots, K\}$ and $\tilde{t} \in g$-th epoch.

We study the event that player j does not select arm $\theta_t(j)$ at time t under two scenarios: (i) $A_j(t) \neq \Omega_*^j(t)$, and (ii) $A_j(t) = \Omega_*^k(t)$, where $\Omega_*^k(t)$ is the set with k best arms at time t. Then, we have

$$\sum_{t=1}^{T}\mathbf{1}\{\varphi_t(j)\neq\theta_t(j),\, t\notin\hat{\mathcal{T}}\}\leq\sum_{t=1}^{T}\mathbf{1}\{A_j(t)\neq\Omega_*^j(t),\, t\notin\hat{\mathcal{T}}\}$$
$$+\sum_{t=1}^{T}\mathbf{1}\{\varphi_t(j)\neq\theta_t(j), A_j(t)=\Omega_*^k(t),\, t\notin\hat{\mathcal{T}}\}. \tag{4.17}$$

Note that unlike RR-SW-UCB#, in SW-DLP, after initialization, $A_k(t)$ is computed every time instead of only at time instants $\{N+\eta M+1\}_{\eta\in\mathbb{Z}_{\geq 0}}$. However, this difference does not change the order of the total number of times that $\Omega_*^k(t)$ is misidentified. Therefore, using (4.16), it follows similarly to the proof of Lemma 4.1 that

$$\sum_{t=1}^{T}\mathbf{1}\{A_j\neq\Omega_*^j(t)\}\in O\big(T^{\frac{1+\nu}{2}}\ln T\big). \tag{4.18}$$

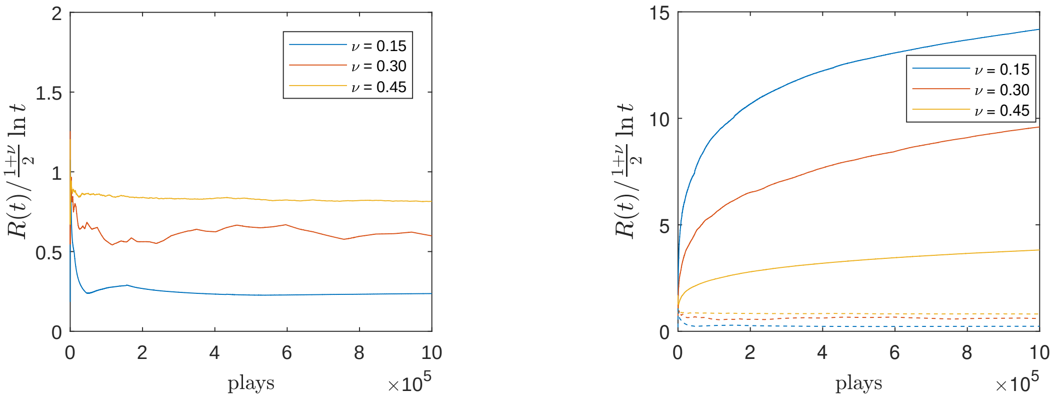
The Chernoff-Hoefding inequality is symmetric about the estimated mean and the upper tail bound is identical to the lower tail bound. Hence, the second term on the right-hand side of inequality (4.17) that involves selecting $\varphi_t(j)$ using lower confidence bounds can be bounded similarly to the first term. Thus, we have

$$\sum_{t=1}^{T}\mathbf{1}\{\varphi_t(j)\neq\theta_t(j),\, A_j=\Omega_*^j(t),\, t\neq\hat{\mathcal{T}}\}\in O\big(T^{\frac{1+\nu}{2}}\ln T\big). \tag{4.19}$$

Substituting (4.18) and (4.19) into (4.17), and substituting (4.17) into (4.15), we conclude that $\hat{\mathcal{N}}_k\in O\big(T^{\frac{1+\nu}{2}}\ln T\big)$.

The number of times the group does not receive a reward from arm $\theta_t(j)$ is upper bounded by the number of times player j does not receive a reward from arm $\theta_t(j)$. Player j does not receive a reward from arm $\theta_t(j)$ if one of the following conditions is true (i) arm $\theta_t(j)$ is not selected by player j, and (ii) arm $\theta_t(j)$ is selected by another player $j'\neq j$. The total number of times either one of these events occurs at any arm $\theta_t(j)$, for all $j\in\{1,\ldots,M\}$, can be upper bounded by $\sum_{k=1}^{M}2\hat{\mathcal{N}}_j$. Since $\hat{\mathcal{N}}_j\in O\big(T^{\frac{1+\nu}{2}}\ln T\big)$ for all $j\in\{1,\ldots,M\}$, we conclude the proof. □

Remark 4.1 (***Comparison of RR-SW-UCB# and SW-DLP***)**.** *In multi-player MAB algorithms, the assignment of a player to a targeted arm is crucial to avoid collisions. In RR-SW-UCB#,*



(a) RR-SW-UCB#

(b) SW-DLP and RR-SW-UCB#

Figure 4.1: Simulation of RR-SW-UCB# and SW-DLP in a piecewise stationary environment.

the indices of arms and the indices of players are employed for this assignment. A round-robin policy ensures all players select M-best arms persistently and accurately estimate the associated mean rewards. While in SW-DLP, such accurate estimation by all players is driven by the lower confidence bound-based assignment of players to the arms. □

4.5 Numerical Illustration

In this section, we present simulation results for RR-SW-UCB# and SW-DLP in abruptly changing environments. In the simulations, we consider a multi-player MAB problem with 6 arms and 3 players. We consider three different values $\{0.15, 0.3, 0.45\}$ of parameter ν that describes the number of breakpoints to show the performance the both algorithms. The breakpoints are introduce at time instants where the next element of sequence $\{\lfloor t^{\nu} \rfloor\}_{t \in \{1,\ldots,T\}}$ is different from current element. We pick them at these time instants to make number of breakpoints $\Upsilon_t \in O(t^{\nu})$ uniformly for all $t \in \{1, \ldots, T\}$. At each break point, the mean rewards at each arm is randomly selected from $\{0.05, 0.22, 0.39, 0.56, 0.73, 0.90\}$. In both algorithms, we select $\lambda = 12.3$.

As shown in Figure. 4.1, with either algorithm, the ratio of the empirical cumulative group regret to the order of $t^{\frac{1+\nu}{2}} \ln t$ is upper bounded by a constant. The dashed lines in Figure. 4.1 (b) are taken directly from (a). The comparison shows that the cumulative regret of RR-SW-UCB# is much lower than SW-DLP. However, if the cost of switching between arms is considered, then the round-robin structure of RR-SW-UCB# would incur significant cost, and in such a scenario SW-DLP might be preferred.

4.6 Summary

We studied the multi-player stochastic MAB problem in abruptly changing environments under a collision model in which a player receives a reward by selecting an arm if it is the only player to select that arm. We designed two novel algorithms, RR-SW-UCB# and SW-DPL to solve this problem. We analyzed these algorithms and characterized their performance in terms of group regret. In particular, we showed that these algorithms incur sublinear expected cumulative regret, i.e., the time average of the regret asymptotically converges to zero. It would be of interest to extend this work to a more general nonstationary environment in which the reward distributions can change at each time step. Another avenue of future research is the extension of these algorithms to the multi-player Markovian MAB problem.

4.7 Bibliographic Remarks

Most of the studies on the multi-player MAB problem deal with a stationary environment. In [22], a lower bound on the expected cumulative group regret for a centralized policy is derived and algorithms that asymptotically achieve this lower bound are designed. Some works assume no communication among players in [4, 5, 23–25], whereas other works allow agents to communicate to improve their arm selection in [26–28].

One of the major generalizations in the multi-player MAB problem is to consider player-dependent rewards, i.e., an arm has different mean rewards for different players [24]. The optimal allocation of the players to arms can be computed using approaches for a famous combinatorial optimization problem known as the assignment problem [96]. To achieve a sublinear regret in a distributed manner, a distributed solution to the assignment problem is required [97]. Assuming collision results in no reward, implicit communication can be generated through collision. In [24], the distributed MAB problem is solved using distributed auction [97] and collision-based implicit communication. The idea of implicit communication is used broadly in different distributed protocols for multi-player MAB problems [98, 99].

More recently, game-theoretic techniques have been used to design fully distributed multi-player MAB algorithms without implicit communication [100]. Specifically, using the payoff dynamics introduced in [101], the authors in [100] design an algorithm that plays, for a sufficiently large portion of time, a strategy profile that optimizes the sum of player-specific mean rewards.

CHAPTER 5

GENERAL NONSTATIONARY BANDITS WITH VARIATION BUDGET

In this chapter, we study a more general non-stationary stochastic MAB problem proposed in [21]. The reward distributions are allowed to either change abruptly like the piece-wise stationary bandits or drift slowly. The nonstationarity of the environment is characterized by the cumulative maximum variation in mean rewards, which subjects to a variation budget.

In order to minimize clutter, we denote the set of arms as $\mathcal{K} := \{1, \ldots, K\}$ and the sequence of time slots as $\mathcal{T} := \{1, \ldots, T\}$. The reward sequence $\{X_t^k\}_{t\in\mathcal{T}}$ for each arm $k \in \mathcal{K}$ is composed of independent samples from potentially time-varying probability distribution function sequence $f_{\mathcal{T}}^k := \{f_t^k(x)\}_{t\in\mathcal{T}}$. We refer to the set $\mathcal{F}_T^{\mathcal{K}} = \{f_{\mathcal{T}}^k \mid k \in \mathcal{K}\}$ containing reward distribution sequences at all arms as the *environment*. Then, the *total variation* of mean rewards in $\mathcal{F}_T^{\mathcal{K}}$ is defined by

$$v\big(\mathcal{F}_T^{\mathcal{K}}\big) := \sum_{t=1}^{T-1} \max_{k\in\mathcal{K}} \left|\mu_{t+1}^k - \mu_t^k\right|, \tag{5.1}$$

which captures the non-stationarity of the environment. We focus on the class of non-stationary environments that have the total variation within a *variation budget* $V_T \geq 0$ which is defined by

$$\mathcal{E}(V_T, T, K) := \big\{\mathcal{F}_T^{\mathcal{K}} \mid v\big(\mathcal{F}_T^{\mathcal{K}}\big) \leq V_T\big\}.$$

The objective is still to design a policy ρ to minimize the regret in a nonstationary environment R_T^ρ defined in (3.1). Note that the performance of a policy ρ differs with different $\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T, T, K)$. For a fixed variation budget V_T and a policy ρ, the *worst-case regret* is the regret with respect to the worst possible choice of environment, i.e.,

$$R_{\text{worst}}^\rho(V_T, T, K) = \sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T,T,K)} R_T^\rho.$$

In this work, we aim at designing policies to minimize the worst-case regret. The optimal worst-case regret achieved by any policy is called the *minimax regret*, and is defined by

$$\inf_{\rho} \sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T,T,K)} R_T^\rho.$$

We study the nonstationary MAB problem under the following two classes of reward distributions:

Assumption 5.1 (Sub-Gaussian reward)**.** *For any $k \in \mathcal{K}$ and any $t \in \mathcal{T}$, distribution $f_t^k(x)$ is* $1/2$ *sub-Gaussian, i.e.,*

$$\forall \lambda \in \mathbb{R} : \mathbb{E}\left[\exp(\lambda(X_t^k - \mu))\right] \leq \exp\left(\frac{\lambda^2}{8}\right).$$

Moreover, for any arm $k \in \mathcal{K}$ and any time $t \in \mathcal{T}$, $\mathbb{E}\left[X_t^k\right] \in [a, a+b]$, where $a \in \mathbb{R}$ and $b > 0$.

Assumption 5.2 (Heavy-tailed reward)**.** *For any arm $k \in \mathcal{K}$ and any time $t \in \mathcal{T}$, $\mathbb{E}\left[(X_t^k)^2\right] \leq 1$.*

5.1 Lower Bound on Minimax Regret in Nonstationary Environment

In this section, we review existing minimax regret lower bounds and minimax policies from literature. These results apply to both sub-Gaussian and heavy-tailed rewards. When $V_T = 0$, the minimax regret lower bound is the same as the one for stochastic stationary bandit (2.1). We show how the minimax regret lower bound for $V_T = 0$ can be extended to establish the minimax regret lower bound for $V_T > 0$. In the later sections, we design a variety of policies that match with the minimax regret lower bound for $V_T > 0$.

In the setting of $V_T > 0$, we recall here the minimax regret lower bound for nonstationary stochastic MAB problems.

Lemma 5.1 (Minimax Lower Bound: $V_T > 0$ [21])**.** *For the non-stationary MAB problem with K arms, time horizon T and variation budget $V_T \in [1/K, T/K]$,*

$$\inf_{\rho} \sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T, T, K)} R_T^{\rho} \geq C(KV_T)^{\frac{1}{3}} T^{\frac{2}{3}},$$

where $C \in \mathbb{R}_{>0}$ is some constant.

To understand this lower bound, consider the following non-stationary environment. The horizon $\mathcal{T}$ is partitioned into epochs of length $\tau = \lceil K^{\frac{1}{3}} (T/V_T)^{\frac{2}{3}} \rceil$. In each epoch, the reward distribution sequences are stationary and all the arms have identical mean rewards except for the unique best arm. Let the gap in the mean be $\Delta = \sqrt{K/\tau}$. The index of the best arm switches at the end of each epoch following some unknown rule. So, the total variation is no greater than $\Delta T/\tau$, which satisfies the variation budget V_T. Besides, for any policy ρ, we know from (2.1) that worst-case regret in each epoch is no less than $C_2\sqrt{K\tau}$. Summing up the regret over all the epochs, minimax regret is lower bounded by $T/\tau \times C_2\sqrt{K\tau}$, which is consistent with Lemma 5.1.

5.2 UCB Algorithms for Sub-Gaussian Nonstationary Stochastic Bandits

In this section, we extend UCB1 and MOSS to design nonstationary UCB policies for scenarios with $V_T > 0$. Three different techniques are employed, namely periodic resetting, sliding observation window and discount factor, to deal with the remembering-forgetting tradeoff. The proposed algorithms are analyzed to provide guarantees on the worst-case regret. We show their performances match closely with the lower bound in Lemma 5.1.

The following notations are used in later discussions. Let $N = \lceil T/\tau \rceil$, for some $\tau \in \{1, \dots, T\}$, and let $\{\mathcal{T}_1, \dots, \mathcal{T}_N\}$ be a partition of time slots $\mathcal{T}$, where each epoch $\mathcal{T}_i$ has length τ except possibly $\mathcal{T}_N$. In particular,

$$\mathcal{T}_i = \Big\{1 + (i-1)\tau\ , \dots,\ \min\left(i\tau, T\right)\Big\},\ i \in \{1, \dots, N\}.$$

Let the maximum mean reward within $\mathcal{T}_i$ be achieved at time $\tau_i \in \mathcal{T}_i$ and arm κ_i, i.e., $\mu_{\tau_i}^{\kappa_i} = \max_{t\in\mathcal{T}_i}\ \mu_t^*$. We define the variation within $\mathcal{T}_i$ as

$$v_i := \sum_{t\in\mathcal{T}_i} \max_{k\in\mathcal{K}} \left|\mu_{t+1}^k - \mu_t^k\right|,$$

where we trivially assign $\mu_{T+1}^k = \mu_T^k$ for all $k \in \mathcal{K}$. Let $\mathbf{1}\{\cdot\}$ denote the indicator function and $|\cdot|$ denote the cardinality of the set, if its argument is a set, and the absolute value if its argument is a real number.

5.2.1 Resetting MOSS Algorithm

Periodic resetting is an effective technique to preserve the freshness and authenticity of the information history. It has been employed in [21] to modify Exp3 to design Rexp3 policy for nonstationary

```
Algorithm 5: The R-MOSS Algorithm
  Input   : $V_T \in \mathbb{R}_{\geq 0}$ and $T \in \mathbb{N}$
  Set     : $\tau = \left\lceil K^{\frac{1}{3}} \left(T/V_T\right)^{\frac{2}{3}} \right\rceil$
  Output  : sequence of arm selection
  while $t \leq T$ do
      if mod $(t, \tau) = 0$ then
          Restart the MOSS policy;
```

stochastic MAB problems. We extend this approach to MOSS and propose a nonstationary policy Resetting MOSS (R-MOSS). In R-MOSS, after every τ time slots, the sampling history is erased and MOSS is restarted. The pseudo-code is provided in Algorithm 5 and the performance in terms of the worst-case regret is established below.

Theorem 5.2. *For the sub-Gaussian nonstationary MAB problem with K arms, time horizon T, variation budget $V_T > 0$, and $\tau = \left\lceil K^{\frac{1}{3}} \left(T/V_T\right)^{\frac{2}{3}} \right\rceil$, the worst case regret of R-MOSS satisfies*

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T, T, K)} R_T^{\text{R-MOSS}} \in \mathcal{O}((KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}).$$

Sketch of the proof. Note that one run of MOSS takes place in each epoch. For epoch $\mathcal{T}_i$, define the set of *bad arms* for R-MOSS by

$$\mathcal{B}_i^{\text{R}} := \{k \in \mathcal{K} \mid \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{k} \geq 2v_i\}. \tag{5.2}$$

Notice that for any $t_1, t_2 \in \mathcal{T}_i$,

$$\left|\mu_{t_1}^k - \mu_{t_2}^k\right| \leq v_i, \quad \forall k \in \mathcal{K}. \tag{5.3}$$

Therefore, for any $t \in \mathcal{T}_i$, we have

$$\mu_t^* - \mu_t^{\varphi_t} \leq \mu_{\tau_i}^{\kappa_i} - \mu_t^{\varphi_t} \leq \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t} + v_i.$$

Then, the regret from $\mathcal{T}_i$ can be bounded as the following,

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mu_t^* - \mu_t^{\varphi_t}\right] \leq |\mathcal{T}_i|\, v_i + \mathbb{E}\left[\sum_{t \in \mathcal{T}_i} \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t}\right] \leq 3|\mathcal{T}_i|\, v_i + S_i, \tag{5.4}$$

where $S_i = \mathbb{E}\left[\sum_{t\in\mathcal{T}_i}\sum_{k\in\mathcal{B}_i^{\mathrm{R}}} \mathbf{1}\left\{\varphi_t = k\right\}\left(\mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t} - 2v_i\right)\right]$.

Now, we have decoupled the problem, enabling us to generalize the analysis of MOSS in the stationary environment [70] to bound S_i. We will only specify the generalization steps and skip the details for brevity.

First notice inequality (5.3) indicates that for any $k \in \mathcal{B}_i^{\mathrm{R}}$ and any $t \in \mathcal{T}_i$,

$$\mu_t^{\kappa_i} \geq \mu_{\tau_i}^{\kappa_i} - v_i \text{ and } \mu_t^k \leq \mu_{\tau_i}^k + v_i.$$

So, at any $t \in \mathcal{T}_i$, $\hat{\mu}_{\kappa_i, n_{\kappa_i}(t)}$ concentrate around a value no smaller than $\mu_{\tau_i}^{\kappa_i} - v_i$, and $\hat{\mu}_{k,n_k(t)}$ concentrate around a value no greater than $\mu_{\tau_i}^k + v_i$ for any $k \in B_i^{\mathrm{R}}$. Also $\mu_{\tau_i}^{\kappa_i} - v_i \geq \mu_{\tau_i}^k + v_i$ due to the definition in (5.2).

In the analysis of MOSS in stationary environment [70], the UCB of each suboptimal arm is compared with the best arm and each selection of suboptimal arm k contribute Δ_k in regret. Here, we can apply a similar analysis by comparing the UCB of each arm $k \in B_i^{\mathrm{R}}$ with κ_i and each selection of arm $k \in B_i^{\mathrm{R}}$ contributes $(\mu_{\tau_i}^{\kappa_i} - v_i) - (\mu_{\tau_i}^k + v_i)$ in S_i. Accordingly, we borrow the upper bound in Lemma 2.3 to get $S_i \leq 49\sqrt{K|\mathcal{T}_i|}$.

Substituting the upper bound on S_i into (5.4) and summarizing over all the epochs, we conclude that

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T,T,K)} R_T^{\text{R-MOSS}} \leq 3\tau V_T + \sum_{i=1}^{N} 49\sqrt{K\tau},$$

which implies the theorem. □

The upper bound in Theorem 5.2 is in the same order as the lower bound in Lemma 5.1. So, the worst-case regret for R-MOSS is order optimal.

5.2.2 Sliding-Window MOSS Algorithm

We have shown that periodic resetting coarsely adapts the stationary policy to a nonstationary setting. However, it is inefficient to entirely remove the sampling history at the restarting points and the regret accumulates quickly close to these points. In [20], a sliding observation window

```
Algorithm 6: The SW-MOSS Algorithm
Input   : $V_T \in \mathbb{R}_{>0}, T \in \mathbb{N}$ and $\eta > 1/2$
Set     : $\tau = \left\lceil K^{\frac{1}{3}} \left(T/V_T\right)^{\frac{2}{3}} \right\rceil$
Output  : sequence of arm selection
Pick each arm once.
while $t \leq T$ do
    Compute statistics within $\mathcal{W}_t = \left\{\min(1, t-\tau), \ldots, t-1\right\}$:
        $\hat{\mu}^k_{n_k(t)} = \frac{1}{n_k(t)} \sum_{s \in \mathcal{W}_t} X_s \mathbf{1}\{\varphi_s = k\}, \; n_k(t) = \sum_{s \in \mathcal{W}_t} \mathbf{1}\{\varphi_s = k\}$
    Pick arm $\varphi_t = \arg\max_{k \in \mathcal{K}} \hat{\mu}^k_{n_k(t)} + \sqrt{\eta \frac{\max\left(\ln\left(\frac{\tau}{K n_k(t)}\right), 0\right)}{n_k(t)}}$;
```

is used to erase the outdated information smoothly and more efficiently utilize the information history. The authors proposed the SW-UCB algorithm that intends to solve the MAB problem with piece-wise stationary mean rewards. We show that a similar approach can also deal with the general nonstationary environment with a variation budget. In contrast to SW-UCB, we integrate the sliding window technique with MOSS instead of UCB1 and achieve the order optimal worst-case regret.

Let the sliding observation window at time t be $\mathcal{W}_t := \left\{\min(1, t-\tau), \ldots, t-1\right\}$. Then, the associated mean estimator is given by

$$\hat{\mu}^k_{n_k(t)} = \frac{1}{n_k(t)} \sum_{s \in \mathcal{W}_t} X_s \mathbf{1}\{\varphi_s = k\}, \; n_k(t) = \sum_{s \in \mathcal{W}_t} \mathbf{1}\{\varphi_s = k\}.$$

For each arm $k \in \mathcal{K}$, define the UCB index for SW-MOSS by

$$g^k_t = \hat{\mu}^k_{n_k(t)} + c_{n_k(k)}, \; c_{n_k(t)} = \sqrt{\eta \frac{\max\left(\ln\left(\frac{\tau}{K n_k(t)}\right), 0\right)}{n_k(t)}},$$

where $\eta > 1/2$ is a tunable parameter. With these notations, SW-MOSS is defined in Algorithm 6. To analyze it, we will use the following concentration bound for sub-Gaussian random variables.

Fact 5.1 (Maximal Hoeffding inequality[83]). *Let $X_1, \ldots, X_n$ be a sequence of independent* $1/2$

sub-Gaussian random variables. Define $d_i := X_i - \mu_i$, then for any $\delta > 0$,

$$\mathbb{P}\left(\exists m \in \{1, \ldots, n\} : \sum_{i=1}^{m} d_i \geq \delta\right) \leq \exp\left(-2\delta^2/n\right),$$

$$\textit{and}\ \mathbb{P}\left(\exists m \in \{1, \ldots, n\} : \sum_{i=1}^{m} d_i \leq -\delta\right) \leq \exp\left(-2\delta^2/n\right).$$

At time t, for each arm $k \in \mathcal{K}$ define

$$M_t^k := \frac{1}{n_k(t)} \sum_{s \in \mathcal{W}_t} \mu_s^k \mathbf{1}_{\{\varphi_s = k\}}.$$

Now, we are ready to present concentration bounds for the sliding window empirical mean $\hat{\mu}^k_{n_k(t)}$.

Lemma 5.3. *For any arm $k \in \mathcal{K}$ and any time $t \in \mathcal{T}$, if $\eta > 1/2$, for any $x > 0$ and $l \geq 1$, the probability of event $A := \left\{\hat{\mu}^k_{n_k(t)} + c_{n_k(t)} \leq M_t^k - x, n_k(t) \geq l\right\}$ is no greater than*

$$\frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau x^2} \exp\left(-x^2 l/\eta\right). \tag{5.5}$$

The probability of event $B := \left\{\hat{\mu}^k_{n_k(t)} - c_{n_k(t)} \geq M_t^k + x, n_k(t) \geq l\right\}$ is also upper bounded by (5.5).

Proof. For any $t \in \mathcal{T}$, let u_i^{kt} be the i-th time slot when arm k is selected within $\mathcal{W}_t$ and let $d_i^{kt} = X^k_{u_i^{kt}} - \mu^k_{u_i^{kt}}$. Note that

$$\mathbb{P}(A) \leq \mathbb{P}\left(\exists m \in \{l, \ldots, \tau\} : \frac{1}{m} \sum_{i=1}^{m} d_i^{kt} \leq -x - c_m\right),$$

Let $a = \sqrt{2\eta}$ such that $a > 1$. We now apply a peeling argument [76, Sec 2.2] with geometric grid $a^s l < m \leq a^{s+1} l$ over $\{l, \ldots, \tau\}$. Since c_m is monotonically decreasing in m,

$$\mathbb{P}\left(\exists m \in \{l, \ldots, \tau\} : \frac{1}{m} \sum_{i=1}^{m} d_i^{kt} \leq -x - c_m\right)$$
$$\leq \sum_{s \geq 0} \mathbb{P}\left(\exists m \in [a^s l, a^{s+1} l) : \sum_{i=1}^{m} d_i^{kt} \leq -a^s l \left(x + c_{a^{s+1} l}\right)\right).$$

According to Fact 5.1, the above summand is no greater than

$$
\begin{aligned}
&\sum_{s\geq 0}\mathbb{P}\left(\exists m\in[1,a^{s+1}l):\sum_{i=1}^{m}d_i^{kt}\leq -a^s l\left(x+c_{a^{s+1}l}\right)\right)\\
\leq&\sum_{s\geq 0}\exp\left(-2\frac{a^{2s}l^2}{\lfloor a^{s+1}l\rfloor}\left(x^2+c^2_{a^{s+1}l}\right)\right)\\
\leq&\sum_{s\geq 0}\exp\left(-2a^{s-1}lx^2-\frac{2\eta}{a^2}\ln\left(\frac{\tau}{Ka^{s+1}l}\right)\right)\\
=&\sum_{s\geq 1}\frac{Kla^s}{\tau}\exp\left(-2a^{s-2}lx^2\right).
\end{aligned}
$$

Let $b=2x^2l/a^2$. It follows that

$$
\begin{aligned}
\sum_{s\geq 1}\frac{Kla^s}{\tau}\exp\left(-ba^s\right)\leq&\frac{Kl}{\tau}\int_0^{+\infty}a^{y+1}\exp\big(-ba^y\big)dy\\
=&\frac{Kla}{\tau\ln(a)}\int_1^{+\infty}\exp(-bz)dz\quad\big(\text{where we set } z=a^y\big)\\
=&\frac{Klae^{-b}}{\tau b\ln(a)},
\end{aligned}
$$

which concludes the bound for the probability of event A. By using upper tail bound, similar result exists for event B. □

We now leverage Lemma 5.3 to get an upper bound on the worst-case regret for SW-MOSS.

Theorem 5.4. *For the nonstationary MAB problem with K arms, time horizon T, variation budget $V_T>0$ and $\tau=\left\lceil K^{\frac{1}{3}}\left(T/V_T\right)^{\frac{2}{3}}\right\rceil$, the worst-case regret of SW-MOSS satisfies*

$$
\sup_{\mathcal{F}_T^{\mathcal{K}}\in\mathcal{E}(V_T,T,K)} R_T^{\textit{SW-MOSS}}\in\mathcal{O}((KV_T)^{\frac{1}{3}}T^{\frac{2}{3}}).
$$

Proof. The proof consists of the following five steps.

Step 1: Recall that v_i is the variation within $\mathcal{T}_i$. Here, we trivially assign $\mathcal{T}_0=\emptyset$ and $v_0=0$. Then, for each $i\in\{1,\ldots,N\}$, let

$$
\Delta_i^k:=\mu_{\tau_i}^{\kappa_i}-\mu_{\tau_i}^k-2v_{i-1}-2v_i,\quad\forall k\in\mathcal{K}.
$$

Define the set of bad arms for SW-MOSS in $\mathcal{T}_i$ as

$$
\mathcal{B}_i^{\mathrm{SW}}:=\{k\in\mathcal{K}\mid\Delta_i^k\geq\epsilon\},
$$

where we assign $\epsilon = 4\sqrt{e\eta K/\tau}$.

Step 2: We decouple the regret in this step. For any $t \in \mathcal{T}_i$, since $\left|\mu_t^k - \mu_{\tau_i}^k\right| \leq v_i$ for any $k \in \mathcal{K}$, it satisfies that

$$\begin{aligned}\mu_t^* - \mu_t^{\varphi_t} &\leq \mu_{\tau_i}^{\kappa_i} - \mu_t^{\varphi_t} \leq \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t} + v_i \\ &\leq \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}\right\}(\Delta_i^{\varphi_t} - \epsilon) + 2v_{i-1} + 3v_i + \epsilon.\end{aligned}$$

Then we get the following inequalities,

$$\begin{aligned}\sum_{t\in\mathcal{T}} \mu_t^* - \mu_t^{\varphi_t} \leq &\sum_{i=1}^{N}\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}\right\}(\Delta_i^{\varphi_t} - \epsilon) + 2v_{i-1} + 3v_i + \epsilon \\ \leq &5\tau V_T + T\epsilon + \sum_{i=1}^{N}\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}\right\}(\Delta_i^{\varphi_t} - \epsilon). \qquad (5.6)\end{aligned}$$

To continue, we take a decomposition inspired by the analysis of MOSS in [70] below,

$$\begin{aligned}&\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}\right\}\left(\Delta_i^{\varphi_t} - \epsilon\right) \\ \leq &\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}, g_t^{\kappa_i} > M_t^{\kappa_i} - \frac{\Delta_i^{\varphi_t}}{4}\right\}\Delta_i^{\varphi_t} \qquad (5.7) \\ +&\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}, g_t^{\kappa_i} \leq M_t^{\kappa_i} - \frac{\Delta_i^{\varphi_t}}{4}\right\}\left(\Delta_i^{\varphi_t} - \epsilon\right), \qquad (5.8)\end{aligned}$$

where summands (5.7) describes the regret when arm κ_i is fairly estimated and summand (5.8) quantifies the regret incurred by underestimating arm κ_i.

Step 3: In this step, we bound the expectation of (5.7). Since $g_t^{\varphi_t} \geq g_t^{\kappa_i}$,

$$\begin{aligned}\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}, g_t^{\kappa_i} > M_t^{\kappa_i} - \frac{\Delta_i^{\varphi_t}}{4}\right\}\Delta_i^{\varphi_t} &\leq \sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}, g_t^{\varphi_t} > M_t^{\kappa_i} - \frac{\Delta_i^{\varphi_t}}{4}\right\}\Delta_i^{\varphi_t} \\ &= \sum_{k\in\mathcal{B}_i^{\mathrm{SW}}}\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k, g_t^k > M_t^{\kappa_i} - \frac{\Delta_i^k}{4}\right\}\Delta_i^k. \qquad (5.9)\end{aligned}$$

Notice that for any $t \in \mathcal{T}_{i-1} \cup \mathcal{T}_i$,

$$\left|\mu_t^k - \mu_{\tau_i}^k\right| \leq v_{i-1} + v_i, \quad \forall k \in \mathcal{K}.$$

It indicates that an arm $k \in \mathcal{B}_i^{\mathrm{SW}}$ is at least Δ_i^k worse in mean reward than arm κ_i at any time slot $t \in \mathcal{T}_{i-1} \cup \mathcal{T}_i$. Since $\mathcal{W}_t \subset \mathcal{T}_{i-1} \cup \mathcal{T}_i$, for any $t \in \mathcal{T}_i$

$$M_t^{\kappa_i} - M_t^k \geq \Delta_i^k \geq \epsilon, \quad \forall k \in \mathcal{B}_i^{\mathrm{SW}}.$$

It follows from (5.9) that

$$\sum_{k \in \mathcal{B}_i^{\mathrm{SW}}} \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k, g_t^k > M_t^{\kappa_i} - \frac{\Delta_i^k}{4}\right\} \Delta_i^k \leq \sum_{k \in \mathcal{B}_i^{\mathrm{SW}}} \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k, g_t^k > M_t^k + \frac{3\Delta_i^k}{4}\right\} \Delta_i^k. \tag{5.10}$$

Let t_s^{ik} be the s-th time slot when arm k is selected within $\mathcal{T}_i$. Then, for any $k \in \mathcal{B}_i^{\mathrm{SW}}$,

$$\begin{aligned}
&\sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k, g_t^k > M_t^k + \frac{3\Delta_i^k}{4}\right\} \\
=& \sum_{s \geq 1} \mathbf{1}\left\{g_{t_s^{ik}}^k > M_{t_s^{ik}}^k + \frac{3\Delta_i^k}{4}\right\} \\
\leq& l_i^k + \sum_{s \geq l_i^k + 1} \mathbf{1}\left\{g_{t_s^{ik}}^k > M_{t_s^{ik}}^k + \frac{3\Delta_i^k}{4}\right\},
\end{aligned} \tag{5.11}$$

where we set $l_i^k = \left\lceil \eta \left(\frac{4}{\Delta_i^k}\right)^2 \ln\left(\frac{\tau}{\eta K}\left(\frac{\Delta_i^k}{4}\right)^2\right)\right\rceil$. Since $\Delta_i^k \geq \epsilon$, for $k \in \mathcal{B}_i^{\mathrm{SW}}$, we have

$$l_i^k \geq \left\lceil \eta \left(4/\Delta_i^k\right)^2 \ln\left(\frac{\tau}{\eta K}\left(\epsilon/4\right)^2\right)\right\rceil \geq \eta\left(4/\Delta_i^k\right)^2,$$

where the second inequality follows by substituting $\epsilon = 4\sqrt{e\eta K/\tau}$. Additionally, since $t_1^{ik}, \ldots, t_{s-1}^{ik} \in \mathcal{W}_{t_s^{ik}}$, we get $n_k(t_s^{ik}) \geq s - 1$. Furthermore, since c_m is monotonically decreasing with m,

$$c_{n_k(t_s^k)} \leq c_{l_i^k} \leq \sqrt{\frac{\eta}{l_i^k} \ln\left(\frac{\tau}{\eta K}\left(\frac{\Delta_i^k}{4}\right)^2\right)} \leq \frac{\Delta_i^k}{4},$$

for $s \geq l_i^k + 1$. Therefore, we continue from (5.11) to get

$$l_i^k + \sum_{s \geq l_i^k + 1} \mathbf{1}\left\{g_{t_s^{ik}}^k > M_{t_s^{ik}}^k + \frac{3\Delta_i^k}{4}\right\} \leq l_i^k + \sum_{s \geq l_i^k + 1} \mathbf{1}\left\{g_{t_s^{ik}}^k - 2c_{n_k(t_s^{ik})} > M_{t_s^{ik}}^k + \frac{\Delta_i^k}{4}\right\}.$$

By applying Lemma 5.3, considering $n_k(t_s^{ik}) \geq s-1$,

$$\begin{aligned}
&\sum_{s \geq l_i^k+1} \mathbb{P}\left\{ g_{t_s^{ik}}^k - 2c_{n_k(t_s^{ik})} > M_{t_s^{ik}}^k + \frac{\Delta_i^k}{4} \right\} \\
\leq &\sum_{s \geq l_i^k} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau} \left(\frac{4}{\Delta_i^k} \right)^2 \exp\left(-\frac{s}{\eta} \left(\frac{\Delta_i^k}{4} \right)^2 \right) \\
\leq &\int_{l_i^k-1}^{+\infty} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau} \left(\frac{4}{\Delta_i^k} \right)^2 \exp\left(-\frac{y}{\eta} \left(\frac{\Delta_i^k}{4} \right)^2 \right) dy \\
\leq &\frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \left(\frac{4}{\Delta_i^k} \right)^4 . \qquad (5.12)
\end{aligned}$$

Let $h(x) = 16\eta/x \ln\left(\tau x^2/16\eta K\right)$ which achieves maximum at $4e\sqrt{\eta K/\tau}$. Combining (5.12), (5.11), (5.10), and (5.9), we obtain

$$\begin{aligned}
\mathbb{E}\left[(5.7)\right] &\leq \sum_{k \in \mathcal{B}_i} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \frac{256}{\left(\Delta_i^k\right)^3} + l_i^k \Delta_i^k \\
&\leq \sum_{k \in \mathcal{B}_i} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \frac{256}{\left(\Delta_i^k\right)^3} + h(\Delta_i^k) + \Delta_i^k \\
&\leq \sum_{k \in \mathcal{B}_i} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \frac{256}{\epsilon^3} + h\left(4e\sqrt{\eta K/\tau}\right) + b \\
&\leq \left(\frac{2.6\eta}{\ln(2\eta)} + 3\sqrt{\eta} \right) \sqrt{K\tau} + Kb.
\end{aligned}$$

Step 4: In this step, we bound expectation of (5.8). When event $\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{SW}}, g_t^{\kappa_i} \leq M_t^{\kappa_i} - \Delta_i^{\varphi_t}/4\right\}$ happens, we know

$$\Delta_i^{\varphi_t} \leq 4M_t^{\kappa_i} - 4g_t^{\kappa_i} \text{ and } g_t^{\kappa_i} \leq M_t^{\kappa_i} - \frac{\epsilon}{4}.$$

Thus, we have

$$\begin{aligned}
&\mathbf{1}\left\{ \varphi_t \in \mathcal{B}_i^{\mathrm{SW}}, g_t^{\kappa_i} \leq M_t^{\kappa_i} - \frac{\Delta_i^{\varphi_t}}{4} \right\} \left(\Delta_i^{\varphi_t} - \epsilon \right) \\
\leq &\mathbf{1}\left\{ g_t^{\kappa_i} \leq M_t^{\kappa_i} - \frac{\epsilon}{4} \right\} \times \left(4M_t^{\kappa_i} - 4g_t^{\kappa_i} - \epsilon \right) := Y.
\end{aligned}$$

Since Y is a nonnegative random variable, its expectation can be computed involving only its cumulative density function:

$$\begin{aligned}\mathbb{E}\ [Y] &= \int_0^{+\infty} \mathbb{P}\left(Y > x\right) dx \\ &\leq \int_0^{+\infty} \mathbb{P}\Big(4M_t^{\kappa_i} - 4g_t^{\kappa_i} - \epsilon \geq x\Big) dx \\ &= \int_\epsilon^{+\infty} \mathbb{P}\Big(4M_t^{\kappa_i} - 4g_t^{\kappa_i} > x\Big) dx \\ &\leq \int_\epsilon^{+\infty} \frac{16(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau x^2} dx = \frac{16(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau\epsilon}.\end{aligned}$$

Hence, $\mathbb{E}\left[(5.8)\right] \leq 16(2\eta)^{\frac{3}{2}} K |\mathcal{T}_i| \,/\big(\ln(2\eta)\tau\epsilon\big)$.

Step 5: With bounds on $\mathbb{E}\ [(5.7)]$ and $\mathbb{E}\left[(5.8)\right]$ from previous steps,

$$\mathbb{E}\left[(5.6)\right] \leq 5\tau V_T + T\epsilon + N\left(\frac{2.6\eta}{\ln(2\eta)} + 3\sqrt{\eta}\right)\sqrt{K\tau} + NKb + \frac{16(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{KT}{\tau\epsilon} \leq C(KV_T)^{\frac{1}{3}} T^{\frac{2}{3}},$$

for some constant C, which concludes the proof. □

We have shown that SW-MOSS also enjoys order optimal worst-case regret. One drawback of the sliding window method is that all sampling history within the observation window needs to be stored. Since window size is selected to be $\tau = \left\lceil K^{\frac{1}{3}}(T/V_T)^{\frac{2}{3}}\right\rceil$, large memory is needed for large horizon length T. The next policy resolves this problem.

5.2.3 Discounted UCB Algorithm

The discount factor is widely used in estimators to forget old information and put more attention on recent information. In [20], such an estimation is used together with UCB1 to solve the piecewise stationary MAB problem, and the policy designed is called Discounted UCB (D-UCB). Here, we tune D-UCB to work in the nonstationary environment with variation budget V_T. Specifically, the mean estimator used is discounted empirical average given by

$$\hat{\mu}_{\gamma,t}^k = \frac{1}{n_{\gamma,t}^k} \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} X_s, \quad n_{\gamma,t}^k = \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\},$$

```
Algorithm 7: The D-UCB Algorithm
  Input   : $V_T \in \mathbb{R}_{>0}$, $T \in \mathbb{N}$ and $\xi > \frac{1}{2}$
  Set     : $\gamma = 1 - K^{-\frac{1}{3}}(T/V_T)^{-\frac{2}{3}}$
  Output  : sequence of arm selection
  for $t \in \{1, \ldots, K\}$ do
      Pick arm $\varphi_t = t$ and set $n^t \leftarrow \gamma^{K-t}$ and $\hat{\mu}^t \leftarrow X_t^t$;
  while $t \leq T$ do
      Pick arm $\varphi_t = \arg\max_{k \in \mathcal{K}} \hat{\mu}^k + 2\sqrt{\frac{\xi \ln(\tau)}{n^k}}$;
      For each arm $k \in \mathcal{K}$, set $n^k \leftarrow \gamma n^k$;
      Set $n^{\varphi_t} \leftarrow n^{\varphi_t} + 1$ & $\hat{\mu}^{\varphi_t} \leftarrow \hat{\mu}^{\varphi_t} + \frac{1}{n^{\varphi_t}}(X_t^{\varphi_t} - \bar{X}^{\varphi_t})$;
```

where $\gamma = 1 - K^{-\frac{1}{3}}(T/V_T)^{-\frac{2}{3}}$ is the discount factor. Besides, the UCB is designed as $g_t^k = \hat{\mu}_t^k + 2c_t^k$, where $c_{\gamma,t}^k = \sqrt{\xi \ln(\tau)/n_{\gamma,t}^k}$ for some constant $\xi > 1/2$. The pseudo code for D-UCB is reproduced in Algorithm 7. It can be noticed that the memory size is only related to the number of arms, so D-UCB requires small memory.

To proceed the analysis, we review the concentration inequality for discounted empirical average, which is an extension of Chernoff-Hoeffding bound. Let

$$M_{\gamma,t}^k := \frac{1}{n_{\gamma,t}^k} \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \mu_s^k.$$

Then, the following fact is a corollary of [20, Theorem 18].

Fact 5.2 (A Hoeffding-type inequality for discounted empirical average with a random number of summands)**.** *For any $t \in \mathcal{T}$ and for any $k \in \mathcal{K}$, the probability of event $A = \left\{\hat{\mu}_{\gamma,t}^k - M_{\gamma,t}^k \geq \delta/\sqrt{n_{\gamma,t}^k}\right\}$ is no greater than*

$$\lceil \log_{1+\lambda}(\tau) \rceil \exp\left(-2\delta^2(1 - \lambda^2/16)\right) \tag{5.13}$$

for any $\delta > 0$ and $\lambda > 0$. The probability of event $B = \left\{\hat{\mu}_{\gamma,t}^k - M_{\gamma,t}^k \leq -\delta/\sqrt{n_{\gamma,t}^k}\right\}$ is also upper bounded by (5.13).

Theorem 5.5. *For the nonstationary MAB problem with K arms, time horizon T, variation budget $V_T > 0$, and $\gamma = 1 - K^{-\frac{1}{3}}(T/V_T)^{-\frac{2}{3}}$, if $\xi > 1/2$, the worst case regret of D-UCB satisfies*

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T,T,K)} R_T^{\text{D-UCB}} \leq C \ln(T)(KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}.$$

Proof. We establish the theorem in four steps.

Step 1: In this step, we analyze $\left|\mu_{\gamma,t}^k - M_{\gamma,t}^k\right|$ at some time slot $t \in \mathcal{T}_i$. Let $\tau' = \log_\gamma \left((1-\gamma)\xi \ln(\tau)/b^2\right)$ and take $t - \tau'$ as a dividing point, then we obtain

$$\begin{aligned}\left|\mu_{\tau_i}^k - M_{\gamma,t}^k\right| \leq & \frac{1}{n_{\gamma,t}^k} \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \left|\mu_{\tau_i}^k - \mu_s^k\right| \\ \leq & \frac{1}{n_{\gamma,t}^k} \sum_{s \leq t-\tau'} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \left|\mu_{\tau_i}^k - \mu_s^k\right| \qquad (5.14) \\ + & \frac{1}{n_{\gamma,t}^k} \sum_{s \geq t-\tau'}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \left|\mu_{\tau_i}^k - \mu_s^k\right|. \qquad (5.15)\end{aligned}$$

Since $\mu_t^k \in [a, a+b]$ for all $t \in \mathcal{T}$, we have (5.14) $\leq b$. Also,

$$(5.14) \leq \frac{1}{n_{\gamma,t}^k} \sum_{s \leq t-\tau'} b\gamma^{t-s} \leq \frac{b\gamma^{\tau'}}{(1-\gamma)n_{\gamma,t}^k} = \frac{\xi \ln(\tau)}{bn_{\gamma,t}^k}.$$

Accordingly, we get

$$(5.14) \leq \min\left(b, \frac{\xi \ln(\tau)}{bn_{\gamma,t}^k}\right) \leq \sqrt{\frac{\xi \ln(\tau)}{n_{\gamma,t}^k}}.$$

Furthermore, for any $t \in \mathcal{T}_i$,

$$(5.15) \leq \max_{s \in [t-\tau', t-1]} \left|\mu_{\tau_i}^k - \mu_s^k\right| \leq \sum_{j=i-n'}^{i} v_j,$$

where $n' = \lceil \tau'/\tau \rceil$ and v_j is the variation within $\mathcal{T}_j$. So we conclude that for any $t \in \mathcal{T}_i$,

$$\left|\mu_{\kappa_i}^k - M_{\gamma,t}^k\right| \leq c_{\gamma,t}^k + \sum_{j=i-n'}^{i} v_j, \quad \forall k \in \mathcal{K}. \qquad (5.16)$$

Step 2: Within partition $\mathcal{T}_i$, let

$$\hat{\Delta}_i^k = \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^k - 2\sum_{j=i-n'}^{i} v_j,$$

and define a subset of bad arms as

$$\mathcal{B}_i^{\mathrm{D}} = \left\{ k \in \mathcal{K} \mid \hat{\Delta}_i^k \geq \epsilon' \right\},$$

where we select $\epsilon' = 4\sqrt{\xi\gamma^{1-\tau} K \ln(\tau)/\tau}$. Since$\left|\mu_t^k - \mu_{\tau_i}^k\right| \leq v_i$ for any $t \in \mathcal{T}_i$ and for any $k \in \mathcal{K}$

$$\begin{aligned}
\sum_{t\in\mathcal{T}} \mu_t^* - \mu_t^{\varphi_t} &\leq \sum_{i=1}^{N} \sum_{t\in\mathcal{T}_i} \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t} + v_i \\
&\leq \tau V_T + \sum_{i=1}^{N} \sum_{t\in\mathcal{T}_i} \left[\mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{D}}\right\} \hat{\Delta}_i^{\varphi_t} + 2 \sum_{j=i-n'}^{i} v_j + \epsilon' \right] \\
&\leq (2n'+3)\tau V_T + N\epsilon'\tau + \sum_{i=1}^{N} \sum_{k\in\mathcal{B}_i^{\mathrm{D}}} \hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k\right\}.
\end{aligned} \tag{5.17}$$

Step 3: In this step, we bound $\mathbb{E}\left[\hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k\right\}\right]$ for an arm $k \in \mathcal{B}_i^{\mathrm{D}}$. Let $t_i^k(l)$ be the l-th time slot arm k is selected within $\mathcal{T}_i$. From arm selection policy, we get $g_t^{\varphi_t} \geq g_t^{\kappa_i}$, which result in

$$\sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k\right\} \leq l_i^k + \sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{ g_t^k \geq g_t^{\kappa_i}, t > t_i^k(l_i^k) \right\}, \tag{5.18}$$

where we pick $l_i^k = \left\lceil 16\xi\gamma^{1-\tau} \ln(\tau)/\left(\hat{\Delta}_i^k\right)^2 \right\rceil$. Note that $g_t^k \geq g_t^{\kappa_i}$ is true means at least one of the followings holds,

$$\hat{\mu}_{\gamma,t}^k \geq M_{\gamma,t}^k + c_{\gamma,t}^k, \tag{5.19}$$

$$\hat{\mu}_{\gamma,t}^{\kappa_i} \leq M_{\gamma,t}^{\kappa_i} - c_{\gamma,t}^{\kappa_i}, \tag{5.20}$$

$$M_{\gamma,t}^{\kappa_i} + c_{\gamma,t}^{\kappa_i} < M_{\gamma,t}^k + 3c_{\gamma,t}^k. \tag{5.21}$$

For any $t \in \mathcal{T}_i$, since every sample before t within $\mathcal{T}_i$ has a weight greater than $\gamma^{\tau-1}$, if $t > t_i^k(l_i^k)$,

$$c_{\gamma,t}^k = \sqrt{\frac{\xi \ln(\tau)}{n_{\gamma,t}^k}} \leq \sqrt{\frac{\xi \ln(\tau)}{\gamma^{\tau-1} l_i^k}} \leq \frac{\hat{\Delta}_i^k}{4}.$$

Combining it with (5.16) yields

$$\begin{aligned}
M_{\gamma,t}^{\kappa_i} - M_{\gamma,t}^k &\geq \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^k - c_{\gamma,t}^{\kappa_i} - c_{\gamma,t}^k - 2\sum_{j=i-n'}^{i} v_j \\
&\geq \hat{\Delta}_i^k - c_{\gamma,t}^{\kappa_i} - c_{\gamma,t}^k \geq 3c_{\gamma,t}^k - c_{\gamma,t}^{\kappa_i},
\end{aligned}$$

which indicates (5.21) is false. As $\xi > 1/2$, we select $\lambda = 4\sqrt{1 - 1/(2\xi)}$ and apply Fact 5.2 to get

$$\mathbb{P}((5.19) \text{ is true}) \leq \left\lceil \log_{1+\lambda}(\tau) \right\rceil \tau^{-2\xi(1-\lambda^2/16)} \leq \frac{\left\lceil \log_{1+\lambda}(\tau) \right\rceil}{\tau}.$$

The probability of (5.20) to be true shares the same bound. Then, it follows from (5.18) that $\mathbb{E}\left[\hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k\right\}\right]$ is upper bounded by

$$\begin{aligned}
&\hat{\Delta}_i^k l_i^k + \hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i} \mathbb{P}\left((5.19) \text{ or } (5.20) \text{ is true}\right)\\
\leq &\frac{16\xi\gamma^{1-\tau}\ln(\tau)}{\hat{\Delta}_i^k} + \hat{\Delta}_i^k + 2\hat{\Delta}_i^k \left\lceil \log_{1+\lambda}(\tau) \right\rceil\\
\leq &\frac{16\xi\gamma^{1-\tau}\ln(\tau)}{\epsilon'} + b + 2b\left\lceil \log_{1+\lambda}(\tau) \right\rceil,
\end{aligned} \tag{5.22}$$

where we use $\epsilon' \leq \hat{\Delta}_i^k \leq b$ in the last step.

Step 4: From (5.17) and (5.22), and plugging in the value of ϵ', an easy computation results in

$$\begin{aligned}
R_T^{\text{D-UCB}} \leq &(2n' + 3)\tau V_T + 8N\sqrt{\xi\gamma^{1-\tau}K\tau\ln(\tau)}\\
&+ 2Nb + 2Nb\log_{1+\lambda}(\tau),
\end{aligned}$$

where the dominating term is $(2n' + 3)\tau V_T$. Considering

$$\tau' = \frac{\ln\left((1-\gamma)\xi\ln(\tau)/b^2\right)}{\ln\gamma} \leq \frac{-\ln\left((1-\gamma)\xi\ln(\tau)/b^2\right)}{1-\gamma},$$

we get $n' \leq C'\ln(T)$ for some constant C'. Hence there exists some absolute constant C such that

$$R_T^{\text{D-UCB}} \leq C\ln(T)(KV_T)^{\frac{1}{3}}T^{\frac{2}{3}}.$$

□

Although the discount factor method requires less memory, there exists an extra factor $\ln(T)$ in the upper bound on the worst-case regret for D-UCB comparing with the minimax regret. This is due to the fact that the discount factor method does not entirely cut off outdated sampling history like periodic resetting or sliding window techniques.

5.3 UCB Policies for Heavy-tailed Nonstationary Stochastic MAB Problems

In this section, we propose and analyze UCB algorithms for the non-stationary stochastic MAB problem with heavy-tailed rewards defined in Assumption 5.2. For the stationary heavy-tailed MAB problem, we have shown Robust MOSS in chapter 2 achieve order optimal worst-case regret. We extend it to the nonstationary setting and design resetting robust MOSS algorithm and sliding-window robust MOSS algorithm.

5.3.1 Resetting robust MOSS for the non-stationary heavy-tailed MAB problem

Like R-MOSS, Resetting Robust MOSS (R-RMOSS) restarts Robust MOSS after every τ time slots. For a stationary heavy-tailed MAB problem, it has been shown in theorem 2.10 that the worst-case regret of Robust MOSS belongs to $O(\sqrt{KT})$. This result along with an analysis similar to the analysis for R-MOSS in Theorem 5.2 yield the following theorem for R-RMOSS. For brevity, we skip the proof.

Theorem 5.6. *For the nonstationary heavy-tailed MAB problem with K arms, horizon T, variation budget $V_T > 0$ and $\tau = \left\lceil K^{\frac{1}{3}} \left(T/V_T\right)^{\frac{2}{3}} \right\rceil$, if $\psi(2\zeta/a) \geq 2a/\zeta$, the worst-case regret of R-RMOSS satisfies*

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T,T,K)} R_T^{\text{R-RMOSS}} \in O((KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}).$$

5.3.2 SW-RMOSS for the non-stationary heavy-tailed MAB problem

In Sliding-Window Robust MOSS (SW-RMOSS), $n_k(t)$ and $\bar{\mu}_{n_k(t)}$ are computed from the sampling history within $\mathcal{W}_t$, and $c_{n_k(t)} = \sqrt{\ln_+ \left(\frac{\tau}{Kn_k(t)}\right)/n_k(t)}$. To analyze SW-RMOSS, we want to establish a similar property as Lemma 5.3 to bound the probability about an arm being under or over estimated. Toward this end, we need the following properties for truncated random variable.

Lemma 5.7. *Let X be a random variable with expected value μ and $\mathbb{E}[X^2] \le 1$. Let $d := \mathrm{sat}(X,B) - \mathbb{E}[\mathrm{sat}(X,B)]$. Then for any $B > 0$, it satisfies (i) $|d| \le 2B$ (ii) $\mathbb{E}[d^2] \le 1$ (iii) $\left|\mathbb{E}[\mathrm{sat}(X,B)] - \mu\right| \le 1/B$.*

Proof. Property (i) follows immediately from definition of d and property (ii) follows from

$$\mathbb{E}[d^2] \le \mathbb{E}\left[\mathrm{sat}^2(X,B)\right] \le \mathbb{E}\left[X^2\right].$$

To see property (iii), since

$$\mu = \mathbb{E}\left[X\left(\mathbf{1}\{|X| \le B\} + \mathbf{1}\{|X| > B\}\right)\right],$$

one have

$$\left|\mathbb{E}[\mathrm{sat}(X,B)] - \mu\right| \le \mathbb{E}\left[\left(|X| - B\right)\mathbf{1}\{|X| > B\}\right] \le \mathbb{E}\left[|X|\,\mathbf{1}\{|X| > B\}\right] \le \mathbb{E}\left[\frac{X^2}{B}\right].$$

□

Moreover, we will also use a maximal Bennett type inequality as shown in the following.

Lemma 5.8 (Maximal Bennett's inequality [75]). *Let $\{X_i\}_{i\in\{1,\ldots,n\}}$ be a sequence of bounded random variables with support $[-B,B]$, where $B \ge 0$. Suppose that $\mathbb{E}[X_i|X_1,\ldots,X_{i-1}] = \mu_i$ and $\mathrm{Var}[X_i|X_1,\ldots,X_{i-1}] \le v$. Let $S_m = \sum_{i=1}^{m}(X_i - \mu_i)$ for any $m \in \{1,\ldots,n\}$. Then, for any $\delta \ge 0$*

$$\mathbb{P}\left(\exists m \in \{1,\ldots,n\} : S_m \ge \delta\right) \le \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right),$$

$$\mathbb{P}\left(\exists m \in \{1,\ldots,n\} : S_m \le -\delta\right) \le \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right).$$

Now, we are ready to establish a concentration property for saturated sliding window empirical mean.

Lemma 5.9. *For any arm $k \in \{1,\ldots,K\}$ and any $t \in \{K+1,\ldots,T\}$, if $\psi(2\zeta/a) \ge 2a/\zeta$, the probability of either event $A = \left\{g_t^k \le M_t^k - x, n_k(t) \ge l\right\}$ or event $B = \left\{g_t^k - 2c_{n_k(t)} \ge M_t^k + x, n_k(t) \ge l\right\}$, for any $x > 0$ and any $l \ge 1$, is no greater than*

$$\frac{2a}{\beta^2 \ln(a)} \frac{K}{\tau x^2}(\beta x\sqrt{h(l)/a} + 1)\exp\left(-\beta x\sqrt{h(l)/a}\right),$$

where $\beta = \psi\left(2\zeta/a\right)/(2a)$.

Proof. Recall that u_i^{kt} is the i-th time slot when arm k is selected within $\mathcal{W}_t$. Since c_m is a monotonically decreasing in m, $1/B_m = c_{h(m)} \leq c_m$ due to $h(m) \geq m$. Then, it follows from property (iii) in Lemma 5.7 that

$$\begin{aligned}
\mathbb{P}(A) &\leq \mathbb{P}\left(\exists m \in \{l, \ldots, \tau\} : \bar{\mu}_m^k \leq \sum_{i=1}^{m} \frac{\mu^k_{u_i^{kt}}}{m} - (1+\zeta)c_m - x\right) \\
&\leq \mathbb{P}\left(\exists m \in \{l, \ldots, \tau\} : \sum_{i=1}^{m} \frac{\bar{d}_{im}^{kt}}{m} \leq \frac{1}{B_m} - (1+\zeta)c_m - x\right) \\
&\leq \mathbb{P}\left(\exists m \in \{l, \ldots, \tau\} : \frac{1}{m}\sum_{i=1}^{m} \bar{d}_{im}^{kt} \leq -x - \zeta c_m\right), \qquad (5.23)
\end{aligned}$$

where $\bar{d}_{im}^{kt} = \mathrm{sat}\left(X^k_{u_i^{kt}}, B_m\right) - \mathbb{E}\left[\mathrm{sat}\left(X^k_{u_i^{kt}}, B_m\right)\right]$. Recall we select $a > 1$. Again, we apply a peeling argument with geometric grid $a^s \leq m < a^{s+1}$ over time interval $\{l, \ldots, \tau\}$. Let $s_0 = \lfloor \log_a(l) \rfloor$. Since c_m is monotonically decreasing with m, we continue from (5.23) to get

$$\mathbb{P}(A) \leq \sum_{s \geq s_0} \mathbb{P}\left(\exists m \in [a^s, a^{s+1}) : \sum_{i=1}^{m} \bar{d}_{im}^{kt} \leq -a^s \left(x + \zeta c_{a^{s+1}}\right)\right). \qquad (5.24)$$

For all $m \in [a^s, a^{s+1})$, since $B_m = B_{a^s}$, from Lemma 5.7 we know $\left|\bar{d}_{im}^{kt}\right| \leq 2B_{a^s}$ and $\mathbf{Var}\left[\bar{d}_{im}^{kt}\right] \leq 1$. Continuing from (5.24), we apply Maximal Bennett's inequality in Lemma 2.7 to get

$$\begin{aligned}
\mathbb{P}(A) \leq& \sum_{s \geq s_0} \exp\left(-\frac{a^s\left(x + \zeta c_{a^{s+1}}\right)}{2B_{a^s}} \psi\left(\frac{2B_{a^s}}{a}\left(x + \zeta c_{a^{s+1}}\right)\right)\right) \\
&\left(\text{since } \psi(x) \text{ is monotonically increasing}\right) \\
\leq& \sum_{s \geq s_0} \exp\left(-\frac{a^s\left(x + \zeta c_{a^{s+1}}\right)}{2B_{a^s}} \psi\left(\frac{2\zeta}{a} B_{a^s} c_{a^{s+1}}\right)\right) \\
&\left(\text{substituting } c_{a^{s+1}}, B_{a^s} \text{ and using } h(a^s) = a^{s+1}\right) \\
=& \sum_{s \geq s_0+1} \exp\left(-a^s\left(\frac{x}{B_{a^{s-1}}} + \zeta c^2_{a^s}\right)\frac{\psi\left(2\zeta/a\right)}{2a}\right) \\
&\left(\text{since } \zeta\psi(2\zeta/a) \geq 2a\right) \\
\leq& \frac{K}{\tau} \sum_{s \geq s_0+1} a^s \exp\left(-a^s \frac{x}{B_{a^{s-1}}} \frac{\psi\left(2\zeta/a\right)}{2a}\right).
\end{aligned}$$

Let $b = x\psi\left(2\zeta/a\right)/(2a)$. Since $\ln_+(x) \geq 1$ for all $x > 0$,

$$
\begin{aligned}
&\frac{K}{\tau}\sum_{s\geq s_0+1} a^s \exp\left(-a^s \frac{x}{B_{a^{s-1}}}\frac{\psi\left(2\zeta/a\right)}{2a}\right)\\
\leq&\frac{K}{\tau}\sum_{s\geq s_0+1} a^s \exp\left(-b\sqrt{a^s}\right)\\
\leq&\frac{K}{\tau}\int_{s_0+1}^{+\infty} a^y \exp\big(-b\sqrt{a^{y-1}}\big)dy\\
=&\frac{K}{\tau}a\int_{s_0}^{+\infty} a^y \exp\big(-b\sqrt{a^{y}}\big)dy\\
=&\frac{K}{\tau}\frac{2a}{\ln(a)b^2}\int_{b\sqrt{a^{s_0}}}^{+\infty} z\exp\big(-z\big)dz \text{ (where } z = b\sqrt{a^y}\text{)}\\
\leq&\frac{K}{\tau}\frac{2a}{\ln(a)b^2}(b\sqrt{a^{s_0}}+1)\exp(-b\sqrt{a^{s_0}}),
\end{aligned}
$$

which concludes the proof. □

With Lemma 5.9, the upper bound on the worst-case regret for SW-RMOSS in the nonstationary heavy-tailed MAB problem can be analyzed similarly as Theorem 5.4.

Theorem 5.10. *For the nonstationary heavy-tailed MAB problem with K arms, time horizon T, variation budget $V_T > 0$ and $\tau = \left\lceil K^{\frac{1}{3}}\left(T/V_T\right)^{\frac{2}{3}}\right\rceil$, if $\psi(2\zeta/a) \geq 2a/\zeta$, the worst-case regret of SW-RMOSS satisfies*

$$
\sup_{\mathcal{F}_T^{\mathcal{K}}\in\mathcal{E}(V_T,T,K)} R_T^{\text{SW-RMOSS}} \leq C(KV_T)^{\frac{1}{3}}T^{\frac{2}{3}}.
$$

Sketch of the proof. The procedure is similar as the proof of Theorem 5.4. The key difference is due to the nuance between the concentration properties on mean estimator. Neglecting the leading constants, the probability upper bound in Lemma 5.3 has a factor $\exp(-x^2 l/\eta)$ comparing with $(\beta x\sqrt{h(l)/a}+1)\exp\left(-\beta x\sqrt{h(l)/a}\right)$ in Lemma 5.9. Since both factors are no greater than 1, by simply replacing η with $(1+\zeta)^2$ and taking similar calculation in every step except inequality (5.12), comparable bounds that only differs in leading constants can be obtained. Applying Lemma 5.9, we revise the computation of (5.12) as the following,

$$
\begin{aligned}
&\sum_{s\geq l_i^k+1} \mathbb{P}\left\{g_{t_s}^k - 2c_{n_k(t_s)} > M_{t_s}^k + \frac{\Delta_i^k}{4}\right\} \\
\leq &\sum_{s\geq l_i^k} C'\left(\frac{\beta\Delta_i^k}{4}\sqrt{\frac{h(l)}{a}}+1\right)\exp\left(-\frac{\beta\Delta_i^k}{4}\sqrt{\frac{h(l)}{a}}\right) \\
\leq &\int_{l_i^k-1}^{+\infty} C'\left(\frac{\beta\Delta_i^k}{4}\sqrt{\frac{y}{a}}+1\right)\exp\left(-\frac{\beta\Delta_i^k}{4}\sqrt{\frac{y}{a}}\right) dy \\
\leq &\frac{6a}{\beta^2}\frac{2a}{\beta^2\ln(a)}\frac{K}{\tau}\left(\frac{4}{\Delta_i^k}\right)^4.
\end{aligned}
\tag{5.25}
$$

where $C' = 2aK\left(4/\Delta_i^k\right)^2/\left(\beta^2\ln(a)\tau\right)$.The second inequality is due to the fact that $(x+1)\exp(-x)$ is monotonically decreasing in x for $x \in [0,\infty)$ and $h(l) > l$. In the last inequality, we change the lower limits of the integration from $l_i^k - 1$ to 0 since $l_i^k \geq 1$ and plug in the value of C'. Comparing with (5.12), this upper bound only varies in constant multiplier. So is the worst-regret upper bound. □

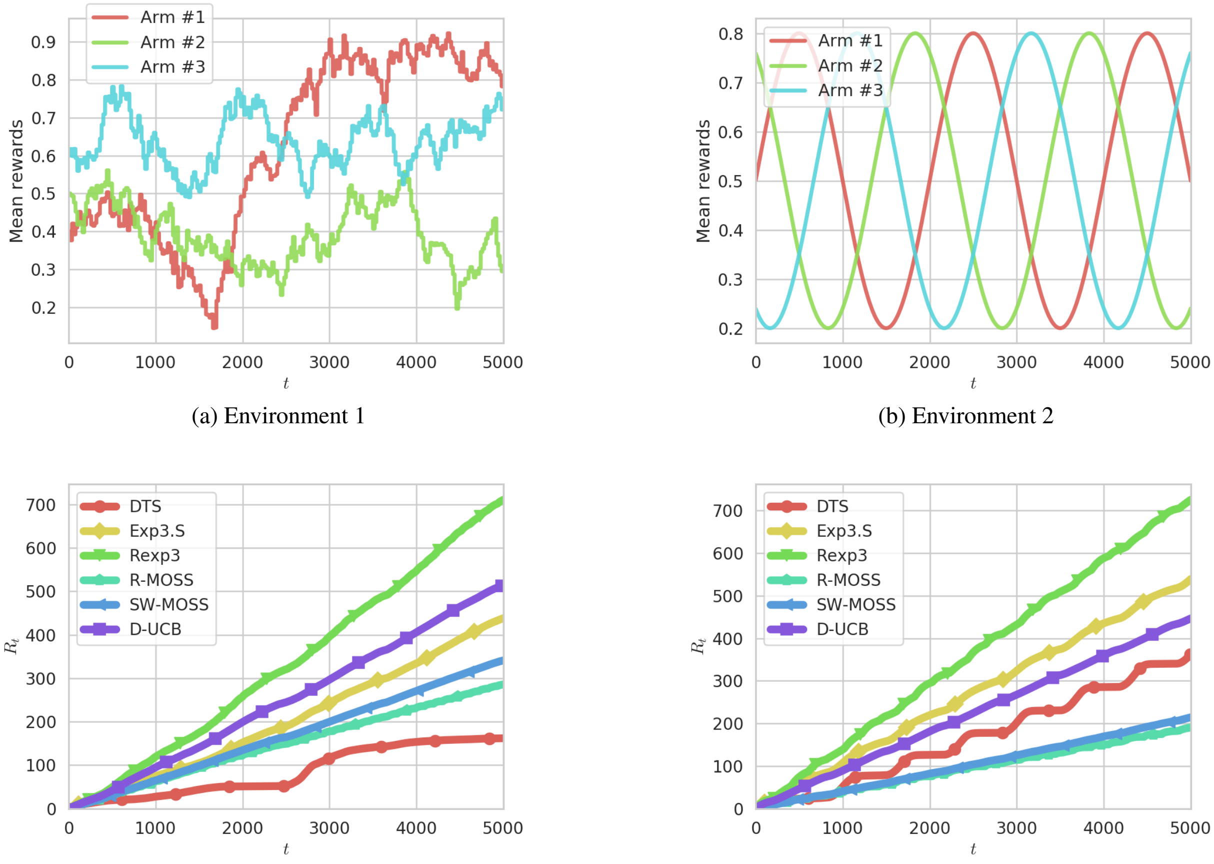
Remark 5.1. *The benefit of the discount factor method is that it is memory-friendly. This advantage is lost if the truncated empirical mean is used. As $n_k(t)$ could both increase and decrease with time, the truncated point could both grow and decline, so all sampling history needs to be recorded. It remains an open problem how to effectively use the discount factor in a nonstationary heavy-tailed MAB problem.*

5.4 Numerical Experiments

We complement the theoretical results in the previous section with two Monte-Carlo experiments. For the light-tailed setting, we compare R-MOSS, SW-MOSS, and D-UCB with other state-of-art policies. For the heavy-tailed setting, we test the robustness of R-RMOSS and SW-RMOSS against both heavy-tailed rewards and nonstationarity. Each result in this section is derived by running designated policies 500 times. And parameter selections for compared policies are strictly coherent with referred literature.

5.4.1 Bernoulli Nonstationay Stochastic MAB Experiment

To evaluate the performance of different policies, we consider two nonstationary environments as shown in Figs. 5.1a and 5.1b, which both have 3 arms with nonstationary Bernoulli rewards. The success probability sequence at each arm is a Brownian motion in environment 1 and a sinusoidal function of time t in environment 2. And the variation budget V_T is 8.09 and 3 respectively.



(a) Environment 1

(b) Environment 2

(c) Regrets for environment 1

(d) Regrets for environment 2

Figure 5.1: Comparison of different policies.

The growths of regret in Figs. 5.1c and 5.1d show that UCB based policies (R-MOSS, SW-MOSS, and D-UCB) maintain their superior performance against adversarial bandit-based policies (Rexp3 and Exp3.S) for stochastic bandits even in nonstationary settings, especially for R-MOSS and SW-MOSS. Besides, DTS outperforms other policies when the best arm does not switch. While each switch of the best arm seems to incur larger regret accumulation for DTS, which results in larger regret compared with SW-MOSS and R-MOSS.

5.4.2 Heavy-tailed Nonstationary Stochastic MAB Experiment

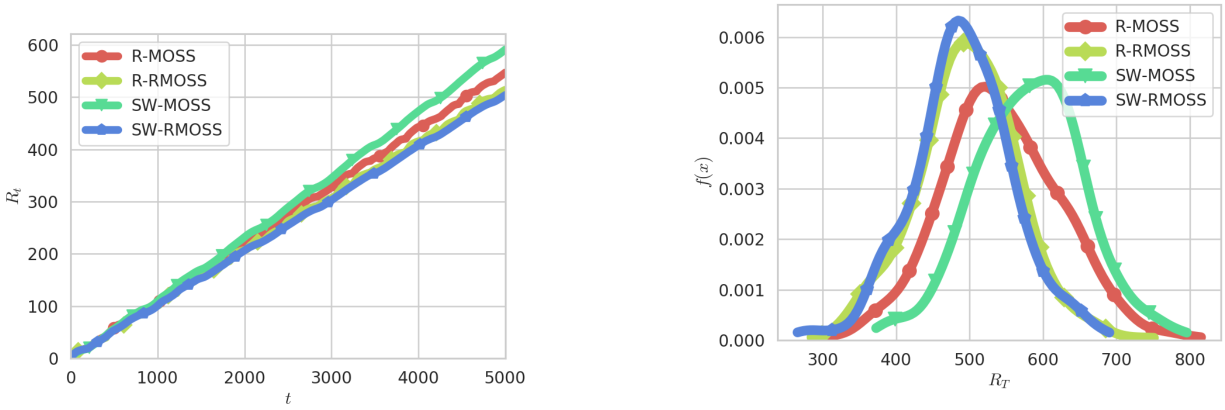
Again we consider the 3-armed bandit problem with sinusoidal mean rewards. In particular, for each arm $k \in \{1, 2, 3\}$,

$$\mu_t^k = 0.3 \sin\left(0.001\pi t + 2k\pi/3\right), \quad t \in \{1, \dots, 5000\}.$$

Thus, the variation budget is 3. Besides, mean reward is contaminated by additive sampling noise ν, where $|\nu|$ is a generalized Pareto random variable and the sign of ν has equal probability to be "+" and "−". So the probability distribution for X_t^k is

$$f_t^k(x) = \frac{1}{2\sigma}\left(1 + \frac{\xi\left|x - \mu_t^k\right|}{\sigma}\right)^{-\frac{1}{\xi}-1} \quad \text{for } x \in (-\infty, +\infty).$$

We select $\xi = 0.4$ and $\sigma = 0.23$ such that Assumption 5.2 is satisfied. We select $a = 1.1$ and $\zeta = 2.2$ for both R-RMOSS and SW-RMOSS such that condition $\psi(2\zeta/a) \geq 2a/\zeta$ is met.



(a) Regret

(b) Histogram of R_T

Figure 5.2: Performances with heavy-tailed rewards.

Figure. 5.2a show RMOSS based polices and slightly outperform MOSS-based polices in heavy-tailed settings. While by comparing the estimated histogram of R_T for different policies in Figure. 5.2b, R-RMOSS and SW-RMOSS have a better consistency and a smaller possibility of a particular realization of the regret deviating significantly from the mean value.

5.5 Summary

We studied the general nonstationary stochastic MAB problem with variation budget and provided three UCB based policies for the problem. Our analysis showed that the proposed policies enjoy the worst-case regret that is within a constant factor of the minimax regret lower bound. Besides, the sub-Gaussian assumption on reward distributions is relaxed to define the nonstationary heavy-tailed MAB problem. We show the order optimal worst-case regret can be maintained by extending the previous policies to robust versions.

There are several possible avenues for future research. In this work, we relied on passive methods to balance the remembering-versus-forgetting tradeoff. The general idea is to keep taking in new information and removing outdated information. Parameter-free active approaches that adaptively detect and react to environment changes are promising alternatives and may result in better experimental performance. Also, extensions from the single decision-maker to distributed multiple decision-makers are of interest. Another possible direction is the nonstationary version of rested and restless bandits.

5.6 Bibliographic Remarks

The adversarial MAB [19] is a paradigmatic nonstationary problem. In this model, the bounded reward sequence at each arm is arbitrary. The performance of a policy is evaluated using the *weak regret*, which is the difference in the cumulated reward of a policy compared with the best single action policy. A $\Omega(\sqrt{KT})$ lower bound on the weak regret and a near-optimal policy Exp3 is also presented in [19]. While being able to capture the nonstationarity, the generality of the reward model in the adversarial MAB makes the investigation of globally optimal policies very challenging.

The nonstationary stochastic MAB can be viewed as a compromise between the stationary stochastic MAB and the adversarial MAB. It maintains the stochastic nature of the reward sequence while allowing some degree of nonstationarity in reward distributions. Instead of the weak regret analyzed in adversarial MAB, a strong notion of regret defined with respect to the best arm at each time step is studied in these problems. As a result, this problem can be studied from two perspectives by extending ideas from adversarial bandits or stochastic bandits.

After formulating the nonstationary stochastic MAB problem in [21], the authors tune the Exp3.S policy for adversarial bandits [19] to achieve a near-optimal worst-case regret in their subsequent work [102]. Discounted Thomson Sampling (DTS) [103] has also been shown to have a good experimental performance within this general framework. However, we are not aware of any analytic regret bounds for the DTS algorithm. The variation budget idea has already been extended to more general problem settings such as nonstationary linear contextual bandits, and the ideas of using periodic resetting, discounting factor, and sliding observation windows have been shown to be applicable therein [104–106]. Nevertheless, to achieve exact order optimal worst-case regret remains unsolved for those generalized problem setups

CHAPTER 6

MULTI-TARGET SEARCH VIA MULTI-FIDELITY GAUSSIAN PROCESSES

The robotic target search problems have a natural connection with MAB problems discussed in the previous section. In particular, the class of robotic search problems in which a robot team searches for a target from a set of view-points (arms), or monitors an environment from a set of viewpoints, maps directly to the MAB problems. In this chapter, we focus on a class of search problems involving the search of an unknown number of targets in a large or continuous space instead of a small number of viewpoints.

We consider a scenario in which an autonomous vehicle equipped with a downward-facing camera operates in a 3D environment, and the task is to search for an unknown number of stationary targets on the 2D floor of the environment. For such a problem, there exists an intrinsic fidelity-vs-coverage trade-off: sensing at a higher altitude provides more global but less accurate information compared with sensing at a lower altitude. To capture this phenomenon, we model the sensing information available at different altitudes from the floor using a multi-fidelity Gaussian process [12]. The key idea to address the fidelity-vs-coverage trade-off is to use the low fidelity information to remove regions unlikely to contain targets. This enables the robot to quickly transit its focus to areas likely to contain targets, thus expedite the search process.

This chapter is a slightly modified version of our published work on multi-target search with multi-fidelity Gaussian process sensing model, and it is reproduced here with the permission of the copyright holder[1]. The proposed multi-target search strategy leverages information-theoretic techniques to efficiently explore the environment, and employ Bayesian techniques to accurately identify targets and construct an occupancy map. The target search accuracy and efficiency are proved with theoretical analysis, and also verified by simulation results.

[1] ©2020 IEEE. Reprinted with permission from [107].

6.1 Multi-target Search Problem Description

We consider an autonomous vehicle that moves in a 3D environment, e.g., an aerial or an underwater vehicle. We assume that the vehicle either moves with unit speed or hovers at a location. The vehicle is tasked with searching for multiple targets on the 2D floor of the environment. Let $D \subset \mathbb{R}^2$ be the area of the floor in which the targets may be present. The vehicle is equipped with a fixed camera that points towards the floor. The vehicle travels across the environment and collects images/videos of the floor (samples) from different sampling points. These sampling points may be located at different altitudes relative to the floor of the environment. We assume that no sample is collected during the movement between sampling points to avoid misleading low-quality sensing information. The collected samples are processed with a computer vision algorithm that outputs a score, which corresponds to the likelihood of a target being present, for each frame. An example of such a computer vision algorithm is the state of art deep neural network YOLOv3 [108]. The score will be used to update the estimate of the sensing output, i.e., the estimated score function $f : D \to [0, 1]$ which will be used to determine the location of the targets. The stochastic model for f is introduced below.

6.1.1 Multi-fidelity Sensing Model

GPs are widely used models for spatially distributed sensing outputs. In [52], a GP is used to model the target detection output of a computer vision algorithm. While target presence is a binary event, the computer vision algorithms such as YOLOv3 yield a score which is a function of the saliency and location of the target in the image. GPs are appropriate models for such score functions. So far in the literature, GPs have been used in the context of single-fidelity measurements. To characterize the inherent fidelity-coverage trade-off in sensing the floor scene by an autonomous vehicle operating in 3D space, we employ a novel multi-fidelity GP model. The two key physical sensing characteristics the model seeks to capture are: (i) there is some information that can only be accessed at lower altitudes, (ii) the sensing outputs are more spatially correlated at higher altitudes, since the fields of view at neighboring locations have higher overlaps in their field of views.

We assume that the vehicle can collect samples of the floor from M possible heights from the floor $z_1 > z_2 > \cdots > z_M$. We refer to these heights as the fidelity level of the measurement, with M (resp. 1) corresponding to the highest (resp. lowest) level of fidelity. Let the score function $g_m : D \to [0,1]$ be defined by the output of the computer vision algorithm for an ideal noise-free image collected at fidelity level $m \in \{1, \ldots, M\}$ with the field of view of the camera centered at $\boldsymbol{x} \in D$. We assume that the score functions for a location $\boldsymbol{x}$ obtained from different altitudes (fidelity levels) are related to each other in an autoregressive manner as follows

$$g^m(\boldsymbol{x}) = a_{m-1} g^{m-1}(\boldsymbol{x}) + b^m(\boldsymbol{x}), \tag{6.1}$$

where a_{m-1} is a scale parameter and b^m is the bias term that captures the information that can be only be accessed at fidelities levels greater than m. Let $f^m(\boldsymbol{x}) = \left(\prod_{i=m}^{M-1} a_i\right) g^m(\boldsymbol{x})$ and $h^m(\boldsymbol{x}) = \left(\prod_{i=m}^{M-1} a_i\right) b^m(\boldsymbol{x})$. Then, equation (6.1) reduces to

$$f^m(\boldsymbol{x}) = f^{m-1}(\boldsymbol{x}) + h^m(\boldsymbol{x}), \tag{6.2}$$

where $f^0(\boldsymbol{x}) = 0$ and $f(\boldsymbol{x}) := f^M(\boldsymbol{x})$ is the score function at the highest fidelity level which we treat as ground truth. We model the influence of systemic errors in sample collection and environmental uncertainty on the output of the computer vision algorithm for an input at fidelity level m through an additive zero mean Gaussian random variable ϵ_m with variance s_m^2, i.e., $\epsilon_m \sim N(0, s_m^2)$. Consequently, the (scaled) score obtained by collecting a sample at location $\boldsymbol{x}$ is a random variable $y = f_m(\boldsymbol{x}) + \epsilon_m$.

We assume that each h_m is a realization of a Gaussian process with a constant mean μ_m and a squared exponential kernel function $k^m(\boldsymbol{x}, \boldsymbol{x}')$ expressed as

$$k^m(\boldsymbol{x}, \boldsymbol{x}') = v_m^2 \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2 l_m^2}\right), \tag{6.3}$$

where l_m is the length scale parameter, and v_m is the variance parameter that satisfies $v_1 > v_2 > \cdots > v_M$. This kernel function describes the spatial correlation of score function at neighboring locations at each fidelity level. Since the fields of view are more overlapped at lower fidelity levels, it results in $l_1 > l_2 > \cdots > l_M$.

We make the following assumptions about the highest-fidelity sample. If the target is not in the field of view at $(\boldsymbol{x}, z_M)$, the mean score of the computer vision algorithm $f(\boldsymbol{x})$ is smaller than a threshold $\mathtt{th}$. If a target is at the center of image collected at $(\boldsymbol{x}, z_M)$, $f(\boldsymbol{x}) \geq \mathtt{th} + \Delta$, for some constant $\Delta > 0$. Here, $1/\Delta$ can be viewed as a measure of detection difficulty that depends both on the quality of the computer vision algorithm and the environment complexity.

6.1.2 Objective of the Multi-target Search Algorithm

Our objective is to design an algorithm for sequentially determining sampling points that lead to expedited detection and localization of targets within desired misclassification rate $\delta \in (0, 1/2)$. In particular, the algorithm should estimate the region containing targets $D_t \subseteq D$ such that (i) $\forall \boldsymbol{x} \in D_t : \mathbb{P}\big(f(\boldsymbol{x}) < \mathtt{th}\big) \leq \delta$ and (ii) $\forall \boldsymbol{x} \in D \setminus D_t : \mathbb{P}\big(f(\boldsymbol{x}) \geq \mathtt{th} + \Delta\big) \leq \delta$. The requirements about both false alarm and mis-detection rate are set by above two conditions.

Let $t(\Delta, \delta)$ be the total (traveling and sampling) time to finish the search task with misclassification rate smaller than δ. Then, the objective of the algorithm is to determine the sequence of sampling points that minimize $t(\Delta, \delta)$.

6.2 Expedited Multi-target Search Algorithm

The proposed Expedited Multi-target Search (EMTS) algorithm is illustrated in Figure. 6.1. It operates using an epoch-based structure. In each epoch, the sampling and fidelity planner computes a set of sampling points and the path planner optimizes a TSP tour going through those points. The vehicle follows the TSP tour to collect measurements at sampling points and the inference algorithm uses these measurements to update the estimate of the score function f. Then, the Bayesian classification uses these estimates to compute an occupancy map of the floor and the region elimination module removes regions with no target with sufficiently high probability from the search space. In the following, we describe each of these modules in detail.

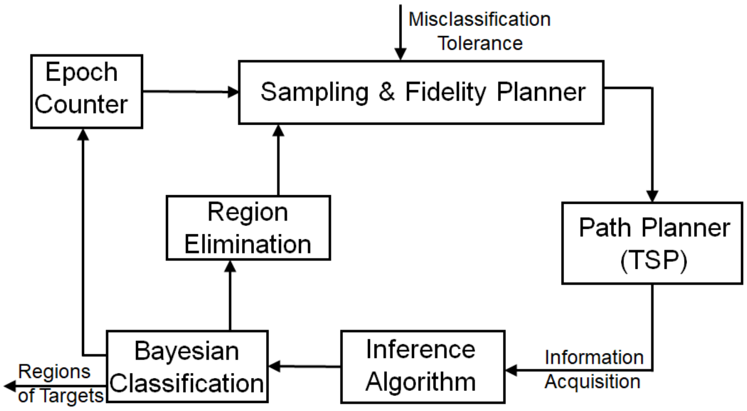


Figure 6.1: Architecture of EMTS.

6.2.1 Inference Algorithm for Multi-fidelity GPs

The Bayesian inference method for multi-fidelity GPs discussed in this section is an extension of the inference procedure in [12] for the case of no sampling noise. Let the set of sampling location-score-fidelity tuples after n observations be $\mathcal{P}_n = \{(\boldsymbol{x}_i, y_i, m_i) \mid i \in \{1, \ldots, n\}\}$. For each fidelity m, define a subset of $\mathcal{P}_n$,

$$P_n^m = \{(\boldsymbol{x}_i, y_i, m_i) \in \mathcal{P}_n \mid m_i = m\},$$

and $|P_n^m|$ denote the cardinality of P_n^m. Recall that $k^i(\boldsymbol{x}, \boldsymbol{x}')$ is the kernel function for the GP h_i at i-th fidelity level. Let $\boldsymbol{K}_0^i(P_n^m, P_n^{m'})$ be a $|P_n^m| \times |P_n^{m'}|$ matrix with entries $k^i(\boldsymbol{x}, \boldsymbol{x}')$, $\boldsymbol{x} \in P_n^m$, $\boldsymbol{x}' \in P_n^{m'}$ and $\boldsymbol{K}_0^i(P_n^m, \boldsymbol{x})$ be a $|P_n^m|$ dimensional vector with entries $k_0^i(\boldsymbol{x}', \boldsymbol{x})$, $\boldsymbol{x}' \in P_n^m$. Let $\boldsymbol{K}$ be a $M \times M$ block matrix with (m, m') block submatrix

$$\boldsymbol{K}_{m,m'} = \sum_{i=1}^{\min(m,m')} \boldsymbol{K}_i\left(P_n^{(m)}, P_n^{(m')}\right).$$

Let $\boldsymbol{k}(\boldsymbol{x})$ be a $|\mathcal{P}_n|$ dimensional vector constructed by concatenating M sub-vectors $\boldsymbol{k}(\boldsymbol{x}) = \left(\boldsymbol{k}^1(\boldsymbol{x}), \ldots, \boldsymbol{k}^M(\boldsymbol{x})\right)$, where

$$\boldsymbol{k}^m(\boldsymbol{x}) = \sum_{i=1}^{m} \boldsymbol{K}_i(P_n^m, \boldsymbol{x}), \quad \forall m \in \{1, \ldots, M\}. \tag{6.4}$$

Denoted by $\boldsymbol{\Theta}$ is the $M \times M$ diagonal matrix with the variance of sampling noise at diagonal entries

$$\boldsymbol{\Theta} = \text{diag}\left\{s_m^2 \boldsymbol{I}_{|P_n^m|}\right\}_{m=\{1,\ldots,M\}}.$$

Let $\boldsymbol{\nu}_n = [\nu_1, \ldots, \nu_n]$ be the a priori mean of the sample $\boldsymbol{y}_n = \left(y_1, \ldots, y_n\right)$. In particular, if y_j is a sample at fidelity m, then $\nu_j = \sum_{i=1}^{m} \mu_i$. The a priori covariance of $\boldsymbol{y}_n$ is $\boldsymbol{K} + \boldsymbol{\Theta}$. In the training process with training dataset $\mathcal{P}_n$, the hyperparameters $\{\mu_m, \nu_m, l_m, s_m\}_{m=1}^{M}$ and $\{a_m\}_{m=1}^{M-1}$ in the multi-fidelity GP can be learned by maximizing a log marginal likelihood function

$$-\frac{1}{2}\log\left(\det\left(2\pi\left(\boldsymbol{K}+\boldsymbol{\Theta}\right)\right)\right) - \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\nu}_n\right)^T\left(\boldsymbol{K}+\boldsymbol{\Theta}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{\nu}_n\right).$$

Such training can be performed using the GP toolbox [109].

Due to the multi-fidelity structure described in (6.1) and (6.2), the prior mean and covariance of f are

$$\mu_0(\boldsymbol{x}) = \sum_{m=1}^{M} \mu_m, \quad k_0(\boldsymbol{x}, \boldsymbol{x}') = \sum_{m=1}^{M} k^m(\boldsymbol{x}, \boldsymbol{x}').$$

When running EMTS with learned hyperparameters, it can be shown that the posterior mean and covariance functions of f after n measurements are

$$\begin{aligned}\mu_n(\boldsymbol{x}) &= \mu_0(\boldsymbol{x}) + \boldsymbol{k}^T(\boldsymbol{x})\left(\boldsymbol{K}+\boldsymbol{\Theta}\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{\nu}_n\right)\\ k_n\left(\boldsymbol{x}, \boldsymbol{x}'\right) &= k_0\left(\boldsymbol{x}, \boldsymbol{x}'\right) - \boldsymbol{k}^T(\boldsymbol{x})\left(\boldsymbol{K}+\boldsymbol{\Theta}\right)^{-1}\boldsymbol{k}(\boldsymbol{x}').\end{aligned} \tag{6.5}$$

Note that the posterior variance $\sigma_n^2(\boldsymbol{x}) = k_n\left(\boldsymbol{x}, \boldsymbol{x}\right)$ is a measure of uncertainty that will be utilized to classify $\boldsymbol{x}$. It should be noted that the measurements collected at different fidelity levels are appropriately scaled in inference (6.5).

6.2.2 Multi-fidelity Sampling & Path Planning

For each epoch j, we seek to design an efficient sampling tour through sampling locations $\{(\boldsymbol{x}_{n_j+1}, z_{n_j+1}), \ldots, (\boldsymbol{x}_{n_{j+1}}, z_{n_{j+1}})\}$ to ensure

$$\max_{\boldsymbol{x}\in D} \sigma_{n_{j+1}}(\boldsymbol{x}) \Big/ \max_{\boldsymbol{x}\in D} \sigma_{n_j}(\boldsymbol{x}) \leq \alpha,$$

where n_j is the number of samples collected before the beginning of the j-th epoch and the selection of uncertainty reduction threshold α is discussed in Section 6.2.3.

Notice that the posterior variance update in (6.5) depends only on the location of the observations $\boldsymbol{y}_n$, but not on the realized value of $\boldsymbol{y}_n$. Therefore, the sequence of sampling location-fidelity tuples can be computed before physically visiting the locations. Such deterministic evolution of the variance has also been leveraged within the context of single-fidelity GP planning to design efficient sampling tours [110].

Sampling Point Selection. The vehicle follows a greedy sampling policy at each fidelity level, i.e., at each sampling round the vehicle selects the most uncertain point as the next sampling point

$$\boldsymbol{x}_n = \underset{\boldsymbol{x} \in D}{\arg\max} \ \ \sigma_{n-1}(\boldsymbol{x}). \tag{6.6}$$

In the information theoretic view [58], the greedy policy is near-optimal in terms of maximizing an appropriate measure of uncertainty reduction.

Fidelity Selection. For each sampling point $\boldsymbol{x}_n$, a fidelity level (or sampling altitude) needs to be assigned. We let the vehicle start at fidelity level 1 and successively visit all fidelity levels from the lowest to the highest. Since sampling f^m is not able to reduce the uncertainty about f introduced by the subsequent bias terms $h^{m+1}, \dots, h^M$, we define the *inaccessible uncertainty* at fidelity level m as $\xi_m = \sum_{i=m+1}^{M} v_i^2$. Accordingly, we define the *accessible uncertainty* about f at fidelity level m by $r_n^m = \max_{\boldsymbol{x} \in D} \sigma_n^2(\boldsymbol{x}) - \xi_m$. The assigned fidelity level to sample point $\boldsymbol{x}_n$ is designed to change from fidelity m to $m+1$ when

$$r_n^m \leq v_{m+1}^2 l_{m+1}^2 / l_m^2.$$

Notice that before the vehicle begins to sample at fidelity level m, $r_n^m \geq v_m^2 \geq v_{m+1}^2 l_{m+1}^2 / l_m^2$, where the second inequality is due to the assumption that $v_m > v_{m+1}$ and $l_m > l_{m+1}$. This ensures that all fidelity levels are visited from the lowest to the highest successively.

Path Planning. Since the order of sampling locations does not influence the eventual posterior mean and variance, the path going through the sampling location can be optimized by computing an approximate TSP tour using packages, such as Concorde [111]. Such a tour-based sampling policy allows for energy and time efficient operation of the vehicle. If all measurements within epoch j are collected at the same fidelity level, the vehicle traverses the TSP tour $\text{TSP}(\boldsymbol{x}_{n_j+1}, \dots, \boldsymbol{x}_{n_{j+1}})$ to collect measurements from sampling points and update posterior distribution of f. Otherwise, a TSP tour is designed at each fidelity level.

6.2.3 Classification and Region Elimination

The classification and elimination of regions follow a confidence-bound-based rule, which has been widely used in pure exploration multi-armed bandit algorithms [112] and robotic source seeking [113]. We extend these ideas to the case of multi-fidelity GP setting.

Conditioned on $\mathcal{P}_n$, the distribution of $f(\boldsymbol{x})$ is Gaussian with mean function $\mu_n(\boldsymbol{x})$ and variance $\sigma_n^2(\boldsymbol{x})$. Let $(L_n(\boldsymbol{x},\varepsilon), U_n(\boldsymbol{x},\varepsilon))$ be the Bayesian confidence interval containing $f(\boldsymbol{x})$ with probability greater than $(1-2\varepsilon)$. Here, the lower confidence bound L_n and upper confidence bound U_n are defined by $L_n(\boldsymbol{x},\varepsilon) = \mu_n(\boldsymbol{x}) - c(\varepsilon)\sigma_n(\boldsymbol{x})$, $U_n(\boldsymbol{x},\varepsilon) = \mu_n(\boldsymbol{x}) + c(\varepsilon)\sigma_n(\boldsymbol{x})$, with $c(\varepsilon) = \sqrt{2\ln\left(1/(2\varepsilon)\right)}$.

Given the desired maximum misclassification rate δ, at the end of epoch j, a location $\boldsymbol{x}$ is classified as *target*, if $L_{n_j}\left(\boldsymbol{x}, \delta/2^j\right) \geq \mathtt{th}$, and is added to D_t; while it is classified as *empty*, if $U_{n_j}\left(\boldsymbol{x}, \delta/2^j\right) < \mathtt{th}$, and is added to the set D_e. Note that the confidence parameter $\varepsilon = \delta/2^j$ defining the lower and upper bounds is decreased exponentially with epochs, and we will show that it ensures a misclassification rate smaller than δ. The locations in the set D_e are removed from sampling space D at the end of each epoch. EMTS is terminated if $\max_{\boldsymbol{x}\in D} 2\sigma_{n_j}(\boldsymbol{x}) \leq \Delta/c(\delta/2^j)$.

The selection of α depends on the balance between the efficiency of the TSP path planer and region elimination. TSP path planer is more effective with smaller α since each exploration tour includes more sample points. While region elimination favors bigger α so that regions not likely to contain targets are removed more frequently.

6.3 An Illustrative Example

In this section, we illustrate EMTS using the Unmanned Underwater Vehicle Simulator [114], which is a ROS package designed for Gazebo robot simulation environment. We integrate it with YOLOv3 [108] for image classification and Concorde solver [111] to compute TSP tours. We use 2 fidelity levels situated at 11m and 5m from the water floor, respectively. In Figure. 6.2, the left figure shows our simulation setup, where an underwater vehicle is equipped with a downward camera and a flashlight to facilitate the searching task in a dark underwater environment. The middle figure

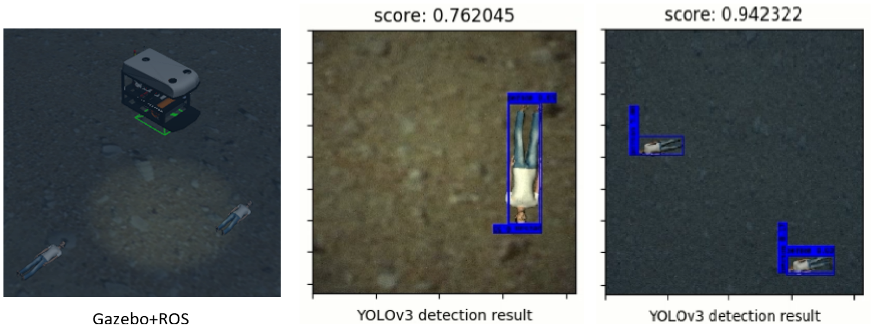


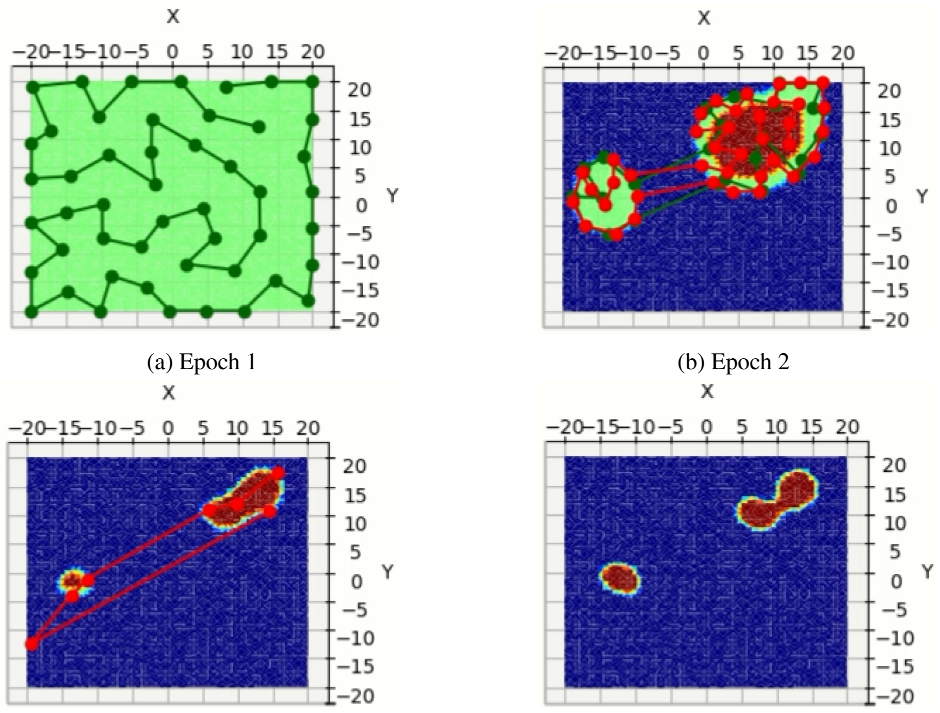
Figure 6.2: Underwater victim search simulation setups.

and right figure in Figure 6.2 are the detection results with YOLOv3 at a high fidelity level and a low fidelity level, respectively. There are 3 victims located at different unknown locations on a $40\text{m} \times 40\text{m}$ water floor. At each sampling point, the vehicle takes 20 images and YOLOv3 returns an average score about the confidence level of the existence of victims in the view.

The first three subplots of Figure. 6.3 show the classification of regions before each epoch, the sampling points selected by the greedy policy and the planned path. Classifications of the environment are represented by 3 colors: red means target exist, blue means no target, and green means uncertain. The dark green points and lines are the planned sampling locations and paths at the low fidelity level and red points and lines are sampling locations and paths at the high fidelity level. At the beginning of epoch 1, all regions are classified as uncertain. After each epoch, the region of targets is narrowed down. The search task is terminated after three epochs. Notice that the vehicle switches to the high fidelity level at epoch 2. The tours at low and high fidelity levels are plotted using two different colors. The vehicles do not sample in blue regions since they have been classified as empty. In the final result, the regions with target are successfully found. A video of the simulation is available online[2].

Figure. 6.4a shows the heat map of posterior variance for the whole region at the end of simulations. It reflects the nature of uncertainty reduction with EMTS, i.e., the posterior variance is low only at areas that likely contain a target. The regions classified as empty have larger posterior variance since they have been eliminated from sampling space in the early phase. This shows that

[2] https://mediaspace.msu.edu/media/EMTS/1_phbul7ui



(a) Epoch 1

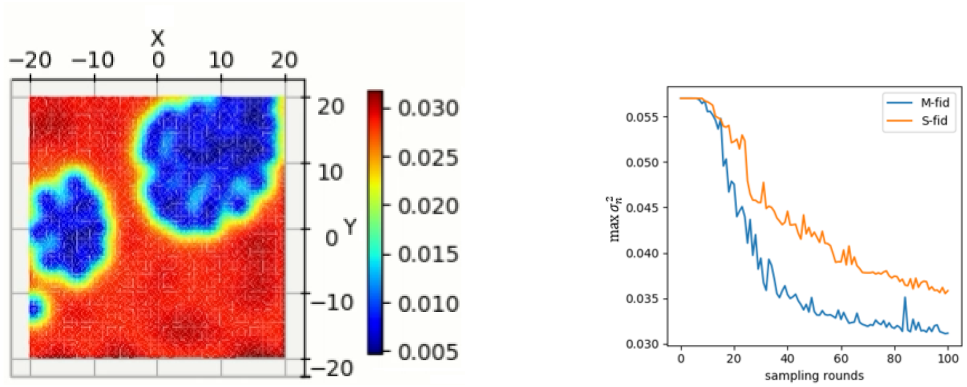
(b) Epoch 2

(c) Epoch 3

(d) Final result

Figure 6.3: Simulation result of EMTS.

EMTS is able to put more focus on areas likely to contain victims. The uncertainty reduction, i.e. the decrease in maximum posterior variance, for multi-fidelity greedy sampling and single-fidelity greedy sampling, are compared in Figure. 6.4b. It shows that greedy multi-fidelity sampling can reduce uncertainty much faster at the beginning stage, which will enable EMTS to eliminate unoccupied regions quickly, and hence, accelerate target search.



(a) Final posterior variance

(b) Convergence of σ_n^2

Figure 6.4: Uncertainty reduction results.

6.4 Analysis of the EMTS Algorithms

In this section, we analyze the modules of the EMTS algorithm and use these analyses to derive an upper bound on the expected detection time for the overall algorithm.

6.4.1 Analysis of the classification algorithm

We first characterize the Bayesian confidence interval for $f(\boldsymbol{x})$, and then use this result to establish that the EMTS algorithm ensures the desired classification accuracy.

Lemma 6.1 (Bayesian confidence interval)**.** *For $f(\boldsymbol{x})\,|\,\mathcal{P}_n \sim N\left(\mu_n(\boldsymbol{x}), \sigma_n^2(\boldsymbol{x})\right)$ and $\varepsilon \in (0, 1/2)$,*

$$\mathbb{P}\left(f(\boldsymbol{x}) \leq L_n(\boldsymbol{x}, \varepsilon)\right) = \mathbb{P}\left(f(\boldsymbol{x}) \geq U_n(\boldsymbol{x}, \varepsilon)\right) \leq \varepsilon.$$

Proof. To normalize $f(\boldsymbol{x})$, let $r = \left(f(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)/\sigma(\boldsymbol{x})$ and $c(\varepsilon) = \sqrt{2\ln\left(1/(2\varepsilon)\right)}$. Now $r \sim N(0, 1)$, and from tail-inequality for standard normal distribution [115]

$$\mathbb{P}\left(r \geq c\right) \leq \frac{1}{2}\exp\left(-\frac{c^2}{2}\right) = \varepsilon,$$

which prove the $\mathbb{P}\left(f(\boldsymbol{x}) \geq U_n(\boldsymbol{x}, \varepsilon)\right) \leq \varepsilon$. Similar result holds for lower confidence bound. □

Theorem 6.2 (Misclassification Rate)**.** *For the classification strategy in the EMTS algorithm, a location $\boldsymbol{x} \in D$ is misclassified with probability at most equal to δ.*

Proof. Consider a location $\boldsymbol{x}$ such that $f(\boldsymbol{x}) \leq \mathtt{th}$, i.e., the true classification of $\boldsymbol{x}$ is *empty*. Since at the end of epoch j, the lower and upper confidence bounds used for classification employ $\varepsilon = \delta/2^j$, we apply a union bound to show the probability of classifying $\boldsymbol{x}$ as a *target* satisfies

$$\sum_{j=1}^{\infty}\mathbb{P}\left(L_{n_j}(\boldsymbol{x}, \delta/2^j) > \mathtt{th}\right) \leq \sum_{j=1}^{\infty}\mathbb{P}\left(L_{n_j}(\boldsymbol{x}, \delta/2^j) > f(\boldsymbol{x})\right).$$

Then, it follows from Lemma 6.1 that the misclassification probability is no greater than $\sum_{j=1}^{\infty}\delta/2^j = \delta$. The case of location $\boldsymbol{x}$ being occupied by a target follows similarly. □

6.4.2 Analysis of the Sampling and Fidelity Planner

We now analyze the information gain and uncertainty reduction properties for our sampling and fidelity planner. We first recall some results for the single fidelity planner and then extend them to the case of the multi-fidelity planner.

Consider a single-fidelity GP f that is sampled with additive Gaussian noise with variance s^2. Let X_n be the set of first n sampling points and let the vector of associated observations be $\boldsymbol{y}_{X_n}$. It is shown in [44, Lemma 5.3] that the mutual information between $\boldsymbol{y}_{X_n}$ and f is

$$I\left(\boldsymbol{y}_{X_n}; f\right) = \frac{1}{2}\sum_{i=1}^{n} \log\left(1 + s^{-2}\sigma_{i-1}^2\left(\boldsymbol{x}_i\right)\right), \tag{6.7}$$

where $\boldsymbol{f}_{X_n}$ is the vector of $f(\boldsymbol{x})$ calculated at points in X_n. Let the maximal mutual information gain with n samples be

$$\gamma_n := \max_{Z \in D:|Z|=n} I\left(\boldsymbol{y}_Z; f\right).$$

Let I_{greedy} be the total mutual information gain using a greedy policy that maximizes the summand in (6.7) at each sampling step. It follows, due to submodularity [116] of $I\left(\boldsymbol{y}_{X_n}; f\right)$, that

$$\left(1 - \frac{1}{e}\right)\gamma_n \leq I_{\text{greedy}}\left(\boldsymbol{y}_{X_n}; f\right) \leq \gamma_n,$$

While giving an exact value of γ_n is difficult, an upper bound on γ_n for squared exponential kernel derived in [44] is presented in the following Lemma 6.3.

Lemma 6.3 (Information gain for squared exp. kernel)**.** *Let a GP f be defined on domain $D \subset \mathbb{R}^2$. If f has squared exponential kernel with length scale l, then the maximum mutual information satisfies*

$$\gamma_n(l) \in O(l^{-2}(\log n)^3).$$

Proof. For a GP defined on $D \in [0,1]^2$ with squared exponential kernel function $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2)$, $\gamma_n \in O((\log n)^3)$ [44]. It is shown in [117] that γ_n scales with the area of D. Thus, if the diameter of D is d, then $\gamma_n \in O\left(d^2(\log n)^3\right)$. Note that having length scale l in kernel function is equivalent to scale D by $1/l$. Accordingly, $\gamma_n \in O\left(d^2 l^{-2}(\log n)^3\right)$. For fixed D, we omit diameter d from the order notation and write $\gamma_n(l) \in O(l^{-2}(\log n)^3)$. □

Lemma 6.3 provides a bound on the mutual information gain at the first fidelity level. For higher fidelity levels, the Gaussian process is composed of the summation of independent GPs. We now establish that the information gained by sampling the sum of GPs is smaller than the information gained by sampling them independently. and then use this result to establish the bound on information gain for multi-fidelity GPs.

Lemma 6.4 (Information gain for sum of GPs)**.** *Let $h_1 \sim GP(\mu_1(\boldsymbol{x}), k_1(\boldsymbol{x}, \boldsymbol{x}'))$ and $h_2 \sim GP(\mu_2(\boldsymbol{x}), k_2(\boldsymbol{x}, \boldsymbol{x}'))$ be independent GPs. Consider a measurement $y = h_1(\boldsymbol{x}) + h_2(\boldsymbol{x}) + \epsilon$ at point $\boldsymbol{x}$, where ϵ is additive measurement noise independent of h_1 and h_2. Let $\boldsymbol{y}_X = \boldsymbol{h}_{1,X} + \boldsymbol{h}_{2,X} + \boldsymbol{\epsilon}$ be the vector of such measurements at sampling points in a set X, where $\boldsymbol{\epsilon}$ is the vector of i.i.d. measurement noise. Then,*

$$I(\boldsymbol{y}_X; h_1 + h_2) \leq I(\boldsymbol{h}_{1,X} + \boldsymbol{\epsilon}; h_1) + I(\boldsymbol{h}_{2,X} + \boldsymbol{\epsilon}; h_2).$$

Proof. The data processing inequality [118, Theorem 2.8.1] indicates

$$I(\boldsymbol{y}_X; h_1 + h_2) \leq I(\boldsymbol{y}_X; h_1, h_2) = I(\boldsymbol{y}_X; h_1) + I(\boldsymbol{y}_X; h_2 \,|\, h_1).$$

Applying the data processing inequality again, we get

$$\begin{aligned} I(\boldsymbol{y}_X; h_1) &\leq I(\boldsymbol{h}_{1,X} + \boldsymbol{\epsilon}, \boldsymbol{h}_{2,X}; h_1) \\ &= I(\boldsymbol{h}_{2,X}; h_1) + I(\boldsymbol{h}_{1,X} + \boldsymbol{\epsilon}; h_1 \,|\, \boldsymbol{h}_{2,X}) \\ &= I(\boldsymbol{h}_{1,X} + \boldsymbol{\epsilon}; h_1), \end{aligned}$$

where in the last step follows due to the independence of h_1, h_2 and $\boldsymbol{\epsilon}$. Similarly, it can be shown

$$I(\boldsymbol{y}_X; h_2 \,|\, h_1) = I(\boldsymbol{h}_{1,X} + \boldsymbol{h}_{2,X} + \boldsymbol{\epsilon}; h_2 \,|\, h_1) = I(\boldsymbol{h}_{2,X} + \boldsymbol{\epsilon}; h_2).$$

This establishes the lemma. □

Let γ_n^m be the maximal mutual information gain at fidelity m. It follows from Lemma 6.4 and the multi-fidelity GP model in (6.2) that $\gamma_n^m \leq \sum_{i=1}^m \gamma_n(l_i)$. Combining this inequality with Lemma 6.3, we obtain the following result.

Corollary 6.5 (Information gain for multi-fidelity GPs)**.** *The maximal mutual information gain at fidelity m satisfies*

$$\gamma_n^m \in O\Big(\sum_{i=1}^{m} l_i^{-2}(\log n)^3\Big).$$

This corollary gives us an insight on the size of γ_n^m at different fidelity levels. It follows that $\gamma_n^{(m)}$ grows faster at higher fidelity levels.

We now derive a bound on the posterior variance for the multi-fidelity GP in terms of the maximum mutual information gain.

Lemma 6.6 (Uncertainty reduction for multi-fidelity GPs)**.** *Let $f \sim GP\left(\mu_0(\boldsymbol{x}), k_0(\boldsymbol{x}, \boldsymbol{x}')\right)$ and $\sigma_0^2(\boldsymbol{x}) \leq \sigma^2$, for each $\boldsymbol{x} \in D$. An additive sampling noise $\epsilon \sim N(0, s^2)$ is incurred every time f is accessed. Under the greedy sampling policy the posterior variance after n sampling rounds satisfies*

$$\max_{\boldsymbol{x}\in D} \sigma_n^2(\boldsymbol{x}) \leq \frac{2\sigma^2}{\log\left(1 + s^{-2}\sigma^2\right)} \frac{\gamma_n}{n}.$$

Proof. For any $\boldsymbol{x} \in D$, $\sigma_n^2(\boldsymbol{x})$ is monotonically non-increasing in n. So we get

$$\max_{\boldsymbol{x}\in D} \sigma_n^2(\boldsymbol{x}) = \sigma_n^2(\boldsymbol{x}_{n+1}) \leq \sigma_{n-1}^2(\boldsymbol{x}_{n+1}) \leq \sigma_{n-1}^2(\boldsymbol{x}_n), \tag{6.8}$$

where the second inequality is due to the fact $\boldsymbol{x}_n = \arg\max_{\boldsymbol{x}\in D} \sigma_{n-1}^2(\boldsymbol{x})$. Again since $\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x}\in D} \sigma_n^2(\boldsymbol{x})$, inequality (6.8) also indicates that $\sigma_{n-1}^2(\boldsymbol{x}_n)$ is monotonically non-increasing. Hence, from (6.7), $\log\left(1 + s^{-2}\sigma_{n-1}^2\left(\boldsymbol{x}_n\right)\right) \leq 2I_{\text{greedy}}\left(\boldsymbol{y}_X; f\right)/n \leq 2\gamma_n/n$. Since $s^2/\log\left(1 + s^2\right)$ is an increasing function on $[0, \infty)$,

$$\sigma_{n-1}^2\left(\boldsymbol{x}_n\right) \leq \frac{\sigma^2}{\log\left(1 + s^{-2}\sigma^2\right)} \log\left(1 + s^{-2}\sigma_{n-1}^2\left(\boldsymbol{x}_n\right)\right).$$

Substituting (6.8) into it, we conclude that

$$\max_{\boldsymbol{x}\in D} \sigma_n^2(\boldsymbol{x}) \leq \sigma_{n-1}^2\left(\boldsymbol{x}_n\right) \leq \frac{2\sigma^2}{\log\left(1 + s^{-2}\sigma^2\right)} \frac{\gamma_n}{n}.$$

□

Lemma 6.6 indicates that the smaller and the more slowly growing γ_n is, the faster $\max_{\boldsymbol{x}\in D} \sigma_n(\boldsymbol{x})$ converges. This result explains our idea of using a multi-fidelity model.

6.4.3 Analysis of Expected Detection Time

We now derive an upper bound on the number of samples needed to classify a location using the EMTS algorithm and then use this result to compute the total sampling and travel time required for classification.

Lemma 6.7 (Sample complexity for uncertainty reduction)**.** *In the autoregressive multi-fidelity model* (6.3)*, if each $h^{(m)}$ has a squared exponential kernel, then*

$$\min\{n \in \mathbb{N} \mid \max_{\boldsymbol{x}\in D} \sigma_n(\boldsymbol{x}) \leq \Delta\} \in O\left(\frac{\sigma_0^2}{\Delta^2}\left(\ln \frac{\sigma_0}{\Delta}\right)^3\right).$$

Proof. It follows from Lemma 6.6 that

$$\frac{n}{\gamma_n} \leq \frac{2\sigma_0^2}{\max_{\boldsymbol{x}\in D} \sigma_n^2(\boldsymbol{x})}.$$

Since v_m, s_m and l_m for all fidelity levels are finite, it follows from Corollary 6.5 that $\gamma_n \in O((\ln n)^3)$. Combining these results, the lemma follows by inspection. □

Lemma 6.8 (Sample complexity for EMTS)**.** *For a given misclassification tolerance δ, let $n(\boldsymbol{x}, \delta)$ be the number of samples required to classify $\boldsymbol{x} \in D$. Then, the expected number of samples satisfies*

$$\mathbb{E}\left[n(\boldsymbol{x}, \delta) \mid \Delta(\boldsymbol{x})\right] \in O\left(\varphi(\Delta(\boldsymbol{x}), \delta)\left(\ln \varphi(\Delta(\boldsymbol{x}), \delta)\right)^3\right),$$

where $\Delta(\boldsymbol{x}) = \left|f(\boldsymbol{x}) - \mathtt{th}\right|$ and $\varphi(\Delta(\boldsymbol{x}), \delta) = \frac{\sigma_0^2}{\Delta^2(\boldsymbol{x})}\ln\left(\frac{3\sigma_0}{\delta\Delta(\boldsymbol{x})}\right)$.

Proof. Since $\delta < 1/2$, function $c(\delta/2^j)\left(3/4\right)^{j+1}$ is monotonically decreasing for $j \geq 2$. We define

$$J = \left\lceil \log_{4/3}\left(\frac{3\sigma_0}{\Delta(\boldsymbol{x})}\sqrt{2\ln\left(\frac{3\sigma_0}{\delta\Delta(\boldsymbol{x})}\right)}\right)\right\rceil + 1.$$

It can be shown that the choice of J ensures, for $j \geq J$,

$$\begin{aligned} U(\boldsymbol{x}, \delta/2^j) - L(\boldsymbol{x}, \delta/2^j) \leq& 2c(\delta/2^j)\left(3/4\right)^{j+1}\sigma_0 \leq 2c(\delta/2^J)\left(3/4\right)^{J+1}\sigma_0 \\ &\leq \frac{\Delta(\boldsymbol{x})}{2}\sqrt{\frac{\alpha\ln\left(\frac{3\sigma}{\delta\Delta(\boldsymbol{x})}\sqrt{2\ln\frac{3\sigma}{\delta\Delta(\boldsymbol{x})}}\right)}{\ln\frac{3\sigma}{\delta\Delta(\boldsymbol{x})}}} < \Delta(\boldsymbol{x}) \end{aligned} \tag{6.9}$$

where $\alpha = \log_{4/3} 2$ and the second inequality is due to the fact $\ln(x \ln(x))/\ln(x) \le (1+e)/e$. For a point $\boldsymbol{x}$ at which $c^*(\boldsymbol{x}) = 1$ and $\Delta(\boldsymbol{x}) > 0$, based on (6.9), the number of sampling rounds to classify $\boldsymbol{x}$ satisfies

$$\begin{aligned}
n(\boldsymbol{x}, \delta) &\le n_J + \sum_{j=J+1}^{\infty} \mathbf{1} L(\boldsymbol{x}, \delta/2^j) < \mathtt{th} \le U(\boldsymbol{x}, \delta/2^j) \\
&\le n_J + \sum_{j=J+1}^{\infty} \mathbf{1} L(\boldsymbol{x}, \delta/2^j) < \mathtt{th} \\
&\le n_J + \sum_{j=J+1}^{\infty} \mathbf{1} U(\boldsymbol{x}, \delta/2^j) < f(\boldsymbol{x}),
\end{aligned}$$

where n_J is the number of samples collected in the first J epochs. Then the expected sampling rounds can be bounded as

$$\begin{aligned}
\bar{n}(\boldsymbol{x}, \delta) &\le n_J + \sum_{j=J+1}^{\infty} \mathbb{P}\left(L(\boldsymbol{x}, \delta/2^j) \ge \mathtt{th}\right) \\
&\le n_J + \sum_{j=J+1}^{\infty} \mathbb{P}\left(L(\boldsymbol{x}, \delta/2^j) \ge \mathtt{th}\right) \\
&\le n_J + \sum_{j=1}^{\infty} \frac{n_j}{2^j}.
\end{aligned}$$

From Lemma 6.7, we has $n_i \in \tilde{O}((16/9)^j)$. Therefore $\sum_{j=1}^{\infty} n_j/2^j$ is finite. So we conclude

$$\bar{n}(\boldsymbol{x}, \delta) \in O\left(\varphi(\Delta(\boldsymbol{x}), \delta) \left(\ln \varphi(\Delta(\boldsymbol{x}), \delta)\right)^3\right).$$

□

Remark 6.1 (Comparison with sample complexity of multiarmed bandits)**.** *Notice that*

$$\mathbb{E}\left[n(\boldsymbol{x}, \delta) \mid \Delta(\boldsymbol{x})\right] \in \tilde{O}\left(\frac{1}{\Delta^2(\boldsymbol{x})}\right)$$

describes the complexity to of classification of $\boldsymbol{x}$, i.e., for a point with $f(\boldsymbol{x})$ close to $\mathtt{th}$ more time is needed. This term is similar to the sampling complexity [73] in a pure-exploration multi-armed bandit problem. This result is based on the assumption that GPs all have squared exponential kernel. For kernels characterizing less correlations, e.g. Matérn kernels, more sampling rounds are expected. □

We now derive an upper bound on detection time for EMTS.

Theorem 6.9 (Target search time for EMTS)**.** *For a given misclassification tolerance δ and detection difficulty measure $1/\Delta$, the target search time satisfies*

$$t(\Delta, \delta) \in O\left(d^2 \varphi(\Delta, \delta) \left(\ln \varphi(\Delta, \delta)\right)^3\right).$$

Proof. Since we assume unit sampling time, the total sampling time is in the same order as $n(\boldsymbol{x}, \delta)$. Then we consider the traveling time spent in order to collected those samples. Since EMTS requires the vehicle to search from low fidelity level to high fidelity level, the total number of altitude switches is no greater than $M - 1$. As presented in [119], for n points in $[0, 1]^2$, the length of the shortest TSP Tour $< 0.984\sqrt{2n} + 11$. Therefore, the expected traveling time belongs to $O\left(d\sqrt{\bar{n}(\boldsymbol{x}, \delta)}\right)$, where d is the diameter of D. Thus, the expected traveling time belongs to $o(\bar{n}(\boldsymbol{x}, \delta))$. Considering both sampling and traveling time, we conclude $t(\Delta, \delta) \in O\left(d^2 \varphi(\Delta, \delta) \left(\ln \varphi(\Delta, \delta)\right)^3\right)$. □

Theorem 6.9 illustrates the efficiency of the EMTS algorithm, we conjecture it to be near-optimal. This upper bound has a natural implication that the target search time increases with the detection difficulty $1/\Delta$ and the desired classification accuracy $1 - \delta$.

6.5 Summary

We studied the autonomous robotic search of an unknown number of targets located at the 2D floor in an unknown and uncertain 3D environment. The novelty of this work lies in using autoregressive multi-fidelity GPs [12, 117] to model the likelihood of the presence of a target at a location, which is computed by a computer vision algorithm using the sample collected at that location at a given altitude. The multi-fidelity GPs sensing model captures the fact that a high altitude (low fidelity) sample provides more global but less accurate information compared with a low altitude (high fidelity) sample. We designed a multi-target search algorithm EMTS that leverages multi-fidelity Gps to capture the fidelity-coverage trade-off, information-theoretic techniques to efficiently explore the environment, and Bayesian techniques to accurately identify targets and construct an occupancy map. With rigorous analysis, we establish formal guarantees of the target detection accuracy and expected detection time.

6.6 Bibliographic Remarks

Autonomous multi-target search requires the autonomous system to quickly and accurately locate multiple targets of interest in an unknown and uncertain environment. Examples include search and rescue missions, mineral exploration, and tracking natural phenomena. To improve the target search efficiency, the trajectory should be designed to balance the explore-exploit tension—the robot should spend more time at target locations while learning target locations. There have been some efforts to address such explore-exploit tension within the context of informative path planning [9, 11, 36, 45–53].

Gaussian processes (GPs) are the most widely used models for capturing spatiotemporal sensing fields in robotics [42, 43]. Informative path planning using such models of the environment has been studied [51, 53, 58, 120–122]. In a broader class of search problems, robot trajectories are designed to maximize the information collected along the way-points while ensuring that the distance traveled is within a prescribed budget. Such informative path planning problems are studied in [54–58]. While GP-based approaches have been used extensively, most of them rely on single-fidelity measurements. Besides, most of these works focus on maximizing the reduction in uncertainty of the estimates instead of the efficiency of the target search.

The multi-target search can also be viewed as a hot-spot identification problem in which, instead of the global maximum of the field, all locations with values greater than a threshold need to be identified. Such problems have been studied in the multiarmed bandit literature [123, 124]; however, we are not aware of any such studies in the GP setting. Furthermore, all these works focus on single fidelity measurements, while we focus on multiple fidelities of measurements. The multi-target search policy in this chapter can be viewed as a combination of informative path planning and the MAB methods in an environment with multi-fidelity sensing information.

CHAPTER 7

ONLINE ESTIMATION AND COVERAGE CONTROL

An intuitive idea to extend the single robot search policy to N robots is to partition the environment into N regions and allocate each robot to one partition. Ideally, the workload needs to be equitably distributed across all regions to maximize time efficiency. The coverage control focuses on such equitable partitioning problems for a team of robots to provide service to a large or continuous environment. The workload is typically referred to as serving demand in coverage control literature. In the coverage problem [13], a particular demand function ϕ is defined over an environment that specifies the degree to which a robot is "needed". The team of agents aims to partition an environment and achieve a configuration that minimizes the coverage cost defined by the sum of the ϕ-weighted distances from every point in the environment to the nearest agent. Intuitively, more robots should concentrate at the regions with higher demands in order to reduce the coverage cost.

This chapter is a slightly modified version of our published work on adaptive coverage control, and it is reproduced here with the permission of the copyright holder.[1] Unlike the classic coverage control with assume demand function ϕ to be known, we model it as a realization of a Gaussian process that can be learned by taking samples.

7.1 Online Estimation and Coverage Problem

We consider a team of N agents tasked with providing coverage to a finite set of points in an environment represented by an undirected graph. The team is required to navigate within the graph to learn an unknown demand function while maintaining near-optimal configuration. In this section, we present the preliminaries of the estimation and coverage problem.

[1] ©2021 IEEE. Reprinted with permission from [125].

7.1.1 Graph Representation of Environment

We consider a discrete environment modeled by an undirected graph $G = (V, E)$, where the vertex set V contains the finite set of points to be covered and the edge set $E \subseteq V \times V$ is the collection of physically adjacent pairs of vertices that can be reached from each other without passing through other vertices. Let the weight map $w : E \to \mathbb{R}_{>0}$ indicate the distance between connected vertices. We assume G is connected. Following the standard definition of weighted undirected graph, a path in G is an ordered sequence of vertices where there exists an edge between consecutive vertices. The distance between vertices v_i and v_j in G, denoted by $d_G(v_i, v_j)$, is defined by the minimum of the sums of the weights in the paths between v_i and v_j.

Suppose there exists an unknown demand function $\phi : V \to \mathbb{R}_{>0}$ that assigns a nonnegative weight to each vertex in G. Intuitively, $\phi(v_i)$ could represent the intensity of signal of interest such as brightness or column of sound. A robot at vertex v_i is capable of measuring $\phi(v_i)$ by collecting a sample $y = \phi(v_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is an additive zero mean Gaussian noise.

7.1.2 Nonparametric Estimation

Let $\boldsymbol{\phi}$ be a vector with the i-th entry $\phi(v_i)$, $i \in \{1, \dots, |V|\}$, where $|\cdot|$ denotes set cardinality. We assume a multivariate Gaussian prior for $\boldsymbol{\phi}$ such that $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1})$, where $\boldsymbol{\mu}_0$ is the mean vector and $\boldsymbol{\Lambda}_0$ is the inverse covariance matrix. Let $n_i(t)$ be the number of samples and $s_i(t)$ be the summation of sampling results from v_i until time t. Then, the posterior distribution of $\boldsymbol{\phi}$ at time t is $\mathcal{N}\big(\boldsymbol{\mu}(t), \boldsymbol{\Lambda}^{-1}(t)\big)$ [126, Chapter 10], where

$$\begin{aligned} \boldsymbol{\Lambda}(t) &= \boldsymbol{\Lambda}_0 + \sum_{i=1}^{|V|} \frac{n_i(t)}{\sigma^2} \boldsymbol{e}_i \boldsymbol{e}_i^{\mathrm{T}} \\ \boldsymbol{\mu}(t) &= \boldsymbol{\Lambda}^{-1}(t) \left(\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{|V|} \boldsymbol{e}_i \frac{s_i(t)}{\sigma^2} \right). \end{aligned} \tag{7.1}$$

Here, $\boldsymbol{e}_i$ is the standard unit vector with i-th entry to be 1.

7.1.3 Voronoi Partition and Coverage Problem

We define the N-partition of graph G as a collection $P = \{P_i\}_{i=1}^N$ of N nonempty subsets of V such that $\cup_{i=1}^N P_i = V$ and $P_i \cap P_j = \emptyset$ for any $i \neq j$. P is said to be connected if the subgraph induced by P_i denoted by $G[P_i]$ is connected for each $i \in N$. $G[P_i]$ being induced subgraph means its vertex set is P_i and its edge set includes all edges in G whose both end vertices are included in P_i.

The configuration of the robot team is a vector of N vertices $\boldsymbol{\eta} \in V^N$ occupied by the robot team, where the i-th entry η_i corresponds to the position of the i-th robot. The i-th robot is tasked to cover vertices in P_i. Then, the coverage cost corresponding to configuration $\boldsymbol{\eta}$ and connected N-partition P can be defined as

$$\mathcal{H}(\boldsymbol{\eta}, P) = \sum_{i=1}^{N} \sum_{v' \in P_i} d_{G[P_i]}(\eta_i, v')\phi(v'). \tag{7.2}$$

In a coverage problem, the objective is to minimize this coverage cost by selecting appropriate configuration $\boldsymbol{\eta}$ and connected N-partition P. However, how to efficiently find the optimal configuration-partition pair in a large graph with arbitrary demand function ϕ remains an open problem. There are two intermediate results about the optimal selection of configuration or partition when the other is fixed [127].

Optimal Partition with Fixed Configuration For a fixed configuration $\boldsymbol{\eta}$ with distinct entries, a optimal connected N-partition P minimizing coverage cost is called Voronoi partition denoted by $\mathcal{V}(\boldsymbol{\eta})$. Formally, for each $P_i \in \mathcal{V}(\boldsymbol{\eta})$ and any $v' \in P_i$,

$$d_G(v', \eta_i) \leq d_G(v', \eta_j), \quad \forall j \in \{1, \ldots, N\}.$$

Optimal Configuration with Fixed Partition For a fixed connected N-partition P, the centroid of the j-th partition $P_j \in P$ is defined by

$$c_i := \arg\min_{v \in P_i} \sum_{v' \in P_i} d_{G[P_i]}(v, v')\phi(v'),$$

and the optimal configuration is to place one robot at the centroid of each $P_i \in P$. We denote the vector of centroid of P by $\boldsymbol{c}(P)$ with c_i as its i-th element.

Building upon the above two properties, the classic Lloyd algorithm iteratively places the robot to the centroid of the current Voronoi partition and computes the new Voronoi partition using the updated configuration. It is known that the robot team eventually converges to a class of partition called centroidal Voronoi partition defined below.

Definition 7.1 (Centroidal Voronoi Partition, [128]). *An N-partition P is a centroidal Voronoi partion of G if P is a Voronoi partition generated by some configuration with distinct entries $\boldsymbol{\eta}$, i.e. $P = \mathcal{V}(\boldsymbol{\eta})$, and $\boldsymbol{c}\left(\mathcal{V}(\boldsymbol{\eta})\right) = \boldsymbol{\eta}$.*

It needs to be noted that an optimal partition and configuration pair minimizing the coverage cost $\mathcal{H}(\boldsymbol{\eta}, P)$ is of the form $(\boldsymbol{\eta}^*, \mathcal{V}(\boldsymbol{\eta}^*))$, where $\boldsymbol{\eta}^*$ has distinct entries and $\mathcal{V}(\boldsymbol{\eta}^*)$ is a centroidal Voronoi partition. A configuration-partition pair $(\boldsymbol{\eta}', \mathcal{V}(\boldsymbol{\eta}'))$ is considered to be an efficient solution to the coverage problem if $\mathcal{V}(\boldsymbol{\eta}')$ is a centroidal Voronoi partition, even though it is possibly suboptimal [128].

7.1.4 Coverage Performance Evaluation

To achieve efficient coverage, the agents need to balance the trade-off between sampling the environment to learn $\boldsymbol{\phi}$ (exploration) and achieving centroidal Voronoi configuration defined using the estimated ϕ (exploitation). To characterize this trade-off, we introduce a notion of coverage regret.

Definition 7.2 (Coverage Regret). *At each time t, let the team configuration be $\boldsymbol{\eta}_t$ and the connected N-partition be P_t. The coverage regret until time T is defined by $\sum_{t=1}^{T} R_t(\phi)$, where $R_t(\phi)$ is the instantaneous coverage regret with respect to demand function ϕ, and is defined by*

$$R_t(\phi) = 2\mathcal{H}(\boldsymbol{\eta}_t, P_t) - \mathcal{H}(\boldsymbol{c}(P_t), P_t) - \mathcal{H}(\boldsymbol{\eta}_t, \mathcal{V}(\boldsymbol{\eta}_t)),$$

which is the sum of two terms $\mathcal{H}(\boldsymbol{\eta}_t, P_t) - \mathcal{H}(\boldsymbol{c}(P_t), P_t)$ and $\mathcal{H}(\boldsymbol{\eta}_t, P_t) - \mathcal{H}(\boldsymbol{\eta}_t, \mathcal{V}(\boldsymbol{\eta}_t))$. The former (resp., latter) term is the regret induced by the deviation of the current configuration (resp., partition) from the optimal configuration (resp., partition) for the current partition (resp.,

Algorithm 8: Deterministic Sequencing of Learning and Coverage (DSLC)

```
Input  : Environment graph $G$, $\boldsymbol{\mu}_0$ , $\boldsymbol{\Lambda}_0$ ;
Set    : $\alpha \in (0,1)$ and $\beta > 1$;
for epoch $j = 1, 2, \ldots$ do
    Exploration phase:
    The robot team sample at vertices in $V$ to make
        $\max_{i \in \{1,\ldots,|V|\}} \sigma_i^2(t) \leq \alpha^j \sigma_0^2.$
    Information propagation phase:
    Each robot agent propagates its sampling result to the team.
    Each robot update estimated demand function $\hat{\phi}$.
    Coverage phase:
    for $t_j = 1, 2, \ldots, \lceil \beta^j \rceil$ do
        Based on $\hat{\phi}$, run pairwise partitioning algorithm.
```

configuration). Accordingly, no regret is incurred at time t if and only if P_t is a centroidal Voronoi N-partition and $\boldsymbol{\eta}_t = \boldsymbol{c}(P_t)$.

There are two sources contributing to the coverage regret. First, the estimation error in the demand function ϕ. Second, the deviation from centroidal Voronoi partition while sampling environment to learn ϕ.

7.2 Deterministic Sequencing of Learning and Coverage Algorithm

In this section, we describe the Deterministic Sequencing of Learning and Coverage (DSLC) algorithm (Algorithm 8). It operates with an epoch-wise structure and each epoch consists of an exploration (learning) phase and an exploitation (coverage) phase. The exploration phase comprises two sub-phases: estimation and information propagation.

7.2.1 Estimation Phase

Let $\sigma_i^2(t)$ be the marginal posterior variance of $\phi(v_i)$ at time t, i.e., the i-th diagonal entry of $\boldsymbol{\Lambda}^{-1}(t)$. Suppose the marginal prior variance $\sigma_i^2(0) \leq \sigma_0^2$, for each i. Within each epoch j, agents first determine the points to be sampled in order to reduce $\max_{i\in\{1,\ldots,|V|\}} \sigma_i^2(t)$ below a threshold $\alpha^j \sigma_0^2$, where $\alpha \in (0,1)$ is a prespecified parameter.

Note that the posterior covariance computed in (7.1) depends only on the number of samples at each vertex, and does not require actual sampling results. Therefore, the sequence of sampling locations can be computed before physically visiting the locations. Leveraging this deterministic evolution of the covariance, we take a greedy sampling policy that repeatedly selects the point v_{i_t} with maximum marginal posterior variance, i.e.,

$$i_t = \operatorname*{arg\,max}_{i\in\{1,\ldots,|V|\}} \sigma_i(t), \tag{7.3}$$

for $t \in \{\underline{t}_j, \ldots, \bar{t}_j\}$, where $\underline{t}_j$ and $\bar{t}_j$ are the starting and ending time of estimation phase in the j-th epoch. It has been shown that the greedy sampling policy is near-optimal in terms of maximizing the mutual information of the sampling results and demand function ϕ [58].

Let the set of points to be sampled during epoch j be X^j and let $X_r^j = X^j \cap P_{\underline{t}_j,r}$ be the set of sampling points that belong to $P_{\underline{t}_j,r}$, the partition assigned to agent r at time $\underline{t}_j$. Each agent r computes a path through the sampling points in X_r^j and collects noisy measurements from those points. The traveling path can be optimized by solving a Traveling Salesperson Problem (TSP).

Remark 7.1. *With $\mathbf{\Lambda}_0$ as the common knowledge, the set of sampling points X^j for each epoch j can be computed independently by each robot following the greedy sampling policy. If the same tie-breaking rule is followed, the computed X^j and the number of samples at each sampling point are the same for all agents.*

7.2.2 Information Propagation Phase

After the estimation phase, sampling results from each agent must be passed to all other agents. There are several mechanisms to accomplish this in a finite number of steps. For example, agents can communicate with their neighboring agents and use flooding algorithms [129] to relay their sampling results to every agent. Alternatively, the agents may be able to send their sampling results to a cloud and receive global estimates after a finite delay. Another possibility for the agents is to use finite time consensus protocols [130] in the distributed inference algorithm discussed in [28].

For any of the above mechanisms, the sampling results from the entire robot team can be propagated to each robot agent in finite time. Then, each agent has an identical posterior distribution $\mathcal{N}\big(\boldsymbol{\mu}(t), \boldsymbol{\Lambda}^{-1}(t)\big)$ of $\boldsymbol{\phi}$, and $\hat{\phi} := \boldsymbol{\mu}(t)$ will be used as the estimate of the demand function.

7.2.3 Coverage Phase

After the estimation and information propagation phases, agents have the same estimate of the demand function $\hat{\phi}$. The coverage phase involves no environmental sampling and its length is designed to grow exponentially with epochs, i.e., the number of time steps in the coverage phase of the j-th epoch is $\lceil \beta^j \rceil$ for some $\beta > 1$. We use a distributed coverage algorithm, proposed in [127], called pairwise partitioning with the estimated demand function $\hat{\phi}$.

In an connected N-partition P, P_i and P_j is said to be adjacent if there exists a vertex pair $v \in P_i$ and $v' \in P_j$ such that there exist an edge in E connecting v and v'. At each time, a random pair of agents (i, j), with P_i and P_j adjacent, compute an optimal pair of vertices (a^*, b^*) within $P_i \cup P_j$ that minimize

$$\sum_{v' \in P_i \cup P_j} \hat{\phi}(v') \min\Big(d_{G[P_i \cup P_j]}(a, v'), d_{G[P_i \cup P_j]}(b, v')\Big).$$

Then, agents i and j move to a^* and b^*. Subsequently, P_i and P_j are updated to

$$P_i \leftarrow \{v \in P_i \cup P_j \mid d_{G[P_i \cup P_j]}(\eta_i, v) \leq d_{G[P_i \cup P_j]}(\eta_j, v)\},$$

$$P_j \leftarrow \{v \in P_i \cup P_j \mid d_{G[P_i \cup P_j]}(\eta_i, v) > d_{G[P_i \cup P_j]}(\eta_j, v)\}.$$

7.3 Analysis of the DSLC Algorithm

In this section, we analyze DSLC to provide a performance guarantee about the expected cumulative coverage regret. To this end, we leverage the information gain from the estimation phase to analyze the convergence rate of uncertainty. Then, we recall the convergence properties of the pairwise partitioning algorithm used in DSLC. Based on these results, we establish the main result of this paper, i.e., an upper bound on the expected cumulative coverage regret.

7.3.1 Mutual Information and Uncertainty Reduction

Let $X^g = (v_{i_1}, \ldots, v_{i_n})$ be a sequence of n vertices selected by the greedy policy. With a slight abuse of notation, we denote the marginal posterior variance of $\phi(v_i)$ after sampling at $v_{i_1} \ldots v_{i_k}$ by $\sigma_i^2(k)$. We now present a bound on the maximal posterior variance after sampling at vertices within X^g. The following Lemma is adapted from Lemma 6.6 to incorporate the discrete environment. Since the proof steps are similar, we skip them for brevity.

Lemma 7.1 (Uncertainty reduction)**.** *Under the greedy sampling policy, the maximum posterior variance after n sampling rounds satisfies*

$$\max_{i \in \{1,\ldots,|V|\}} \sigma_i^2(n) \leq \frac{2\sigma_0^2}{\log\left(1 + \sigma^{-2}\sigma_0^2\right)} \frac{\gamma_n}{n},$$

where γ_n is the maximal mutual information gain that can be achieved with n samples.

Typically, it is hard to characterize γ_n with a general $\mathbf{\Sigma}_0$. Therefore, we assume a squared exponetial kernel for $\boldsymbol{\phi}$.

Assumption 7.1. *Vertices in V lie in a convex and compact set $D \in \mathbb{R}^2$ and the covariance of any pair $\phi(v_i)$ and $\phi(v_j)$ is determined by a squared exponential kernel function*

$$k(\phi(v_i), \phi(v_j)) = \sigma_v^2 \exp\left(-\frac{d_{\mathrm{eu}}^2(v_i, v_j)}{2l^2}\right),$$

where $d_{\mathrm{eu}}(v_i, v_j)$ is the Euclidean distance between v_i and v_j, l is the length scale, and σ_v^2 is the variability parameter.

We now recall an upper bound on γ_n from [44].

Lemma 7.2 (Information gain for squared exp. kernel)**.** *With Assumption 7.1, the maximum mutual information satisfies $\gamma_n \in O((\log|V|\, n)^3)$.*

Remark 7.2. *If the correlation information is ignored, i.e., $\phi(i)$, $i \in \{1, \ldots, |V|\}$ are treated to be independent, it can be seen that $\max_{i\in\{1,\ldots,|V|\}} \sigma_i^2(n) \in O(|V|/n)$ with greedy sampling policy. In contrast, if correlation information is considered, by substituting the result in Lemma 7.2 into Lemma 7.1, $\max_{i\in\{1,\ldots,|V|\}} \sigma_i^2(n) \in O((\log(|V|n))^3/n)$, which shows great advantage about reducing uncertainty when $|V|$ is large (the environment is finely discretized).*

7.3.2 Convergence within Coverage Phase

Before each coverage phase, since the sampling results of each agent are relayed to the entire team, the agents have a consensus estimate of the demand function $\hat{\phi}$. It has been shown in [127] that using the pairwise partitioning algorithm, the N-partition P for the team converges almost surely to a class of near-optimal partitions defined below.

Definition 7.3 (Pairwise-optimal Partition)**.** *A connected N-partition P is pairwise-optimal if for each pair of adjacent regions P_i and P_j,*

$$\sum_{v'\in P_i} d_G(c(P_i), v')\phi(v') + \sum_{v'\in P_j} d_G(c(P_j), v')\phi(v')$$
$$= \min_{a,b\in P_i\cup P_j} \sum_{v'\in P_i\cup P_j} \phi(v') \min\left(d_G(a, v'), d(b, v')\right).$$

It means that, within the induced subgraph generated by any pair of adjacent regions, the 2-partition is optimal. It is proved in [127] that if a connected N-partition P is pairwise-optimal then it is also a centroidal Voronoi partition. The following result on the convergence time of the pairwise partitioning algorithm is established in [127].

Lemma 7.3 (Expected Convergence Time)**.** *Using the pairwise partitioning algorithm, the expected time to converge to a pairwise-optimal N-Partition is finite.*

For each coverage phase, Lemma 7.3 implies that the expected time for the instantaneous regret $R_t(\hat{\phi})$ to converge to 0 is finite.

7.3.3 An Upper Bound on Expected Coverage Regret

We now present the main result for this paper.

Theorem 7.4. *For DSLC and any time horizon T, if Assumption 7.1 holds and $\alpha = \beta^{-2/3}$, then the expected cumulative coverage regret with respect to demand function ϕ satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t(\phi)\right] \in O\big(T^{2/3}(\log(T))^3\big).$$

Proof. We establish the theorem using the following four steps.

Step 1 (Regret from estimation phases): Let the total number of sampling steps before the end of the j-th epoch be s_j. By applying Lemma 7.1, we get

$$s_j \in O((\log(T))^3/\alpha^j).$$

Thus, the coverage regret in the estimation phases until the end of the j-th epoch belongs to $O((\log(T))^3/\alpha^j)$.

Step 2 (Regret from information propagation phases): The sampling information by each robot propagates to all the team members in finite time. Thus, before the end of the j-th epoch, the coverage regret from information propagation phases can be bounded by $c_1 j$ for some constant $c_1 > 0$.

Step 3 (Regret from coverage phases): According to Lemma 7.3, in each coverage phase, the expected time before converging to a pairwise-optimal partition is finite. Thus, before the end of the j-th epoch, the expected coverage regret from converging steps can be upper bounded by $c_2 j$ for some constant $c_2 > 0$.

Also note that the robot team converge to pair-wise optimal partition with respect estimated demand function $\hat{\phi}$ which may deviate from the actual ϕ. Then, the instantaneous coverage regret $R_t(\phi)$ caused by estimation error can be expressed as

$$2\mathcal{H}(\boldsymbol{\eta}_t, P_t) - \mathcal{H}(\boldsymbol{c}(P_t), P_t) - \mathcal{H}(\boldsymbol{\eta}_t \mathcal{V}(\boldsymbol{\eta}_t)) := A_t^{\mathrm{T}}\boldsymbol{\phi},$$

for some $A_t \in \mathbb{R}^{|V|}$. Moreover, the posterior distribution of $R_t(\phi)$ is $\mathcal{N}(A_t^{\mathrm{T}}\boldsymbol{\mu}(t), A_t^{\mathrm{T}}\boldsymbol{\Sigma}(t)A_t)$, where $\boldsymbol{\Sigma}(t) = \boldsymbol{\Lambda}^{-1}(t)$ is the posterior covariance matrix. Since a pairwise-optimal partition P is also a centroidal Voronoi partition and $\hat{\phi} = \boldsymbol{\mu}(t)$, $R_t(\hat{\phi}) = 0$ indicates $A_t^{\mathrm{T}}\boldsymbol{\mu}(t) = 0$. Now, we get $R_t(\phi) \sim \mathcal{N}(0, A_t^{\mathrm{T}}\boldsymbol{\Sigma}(t)A_t)$ and

$$\mathbb{E}\left[R_t(\phi)\right] \leq \mathbb{E}\left[\left\|R_t(\phi)\right\|\right] = \sqrt{\frac{2}{\pi}A_t^{\mathrm{T}}\boldsymbol{\Sigma}(t)A_t}.$$

Note that $A_t^{\mathrm{T}}\boldsymbol{\Sigma}(t)A_t$ is weighed summation of eigenvalues of $\boldsymbol{\Sigma}(t)$. At any time t in the coverage phase of the k-th epoch, $\max_{i\in\{1,\dots,|V|\}}\sigma_i^2(t) \leq \alpha^k\sigma_0^2$, and its follows that the summation of eigenvalues of $\boldsymbol{\Sigma}(t)$ equals $\operatorname{trace}(\boldsymbol{\Sigma}(t)) \leq |V|\,\alpha^k\sigma_0^2$. Thus, we get

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_k^{\mathrm{cov}}} R_t(\phi)\right] \leq c_3(\beta\sqrt{\alpha})^k,$$

for some constant $c_3 > 0$, where $\mathcal{T}_k^{\mathrm{cov}}$ are the time slots in the coverage phase of the k-th epoch and we have used the fact that $|\mathcal{T}_k^{\mathrm{cov}}| = \lceil\beta^k\rceil$.

Step 4 (Summary): Summing up the expected coverage regret from the above steps, the expected cumulative coverage regret at the end of the j-th epoch T_j satisfies

$$\mathbb{E}\left[\sum_{t=1}^{T_j} R_t(\phi)\right] \leq C_1 j + C_2 s_j + \sum_{k=1}^{j} c_3(\beta\sqrt{\alpha})^k, \tag{7.4}$$

where $C_1, C_2 > 0$ are some constants. The theorem statement follows by plugging in $\alpha = \beta^{-2/3}$, using $j \in O(\log T)$ and some simple calculations. □

7.4 Simulation Results

To illustrate the empirical performance of the proposed algorithm, we simulate its execution on a uniform grid graph superimposed on the unit square. We present numerical results which show that DSLC achieves sublinear regret and compare our algorithm to those proposed in [13] and [68].

Motivated by environmental applications, we construct the demand function ϕ over a discrete 21×21 point grid in $[0,1]^2$ by performing kernel density estimation on a subset of the geospatial distribution of Australian wildfires observed by NASA in 2019 [131]. Intuitively, ϕ represents the probability that a wildfire may occur at a particular point of the unit square, and it is used to model the demand for an autonomous sensing agent at that point. The ground truth ϕ is shown on the right in Figure 7.1.

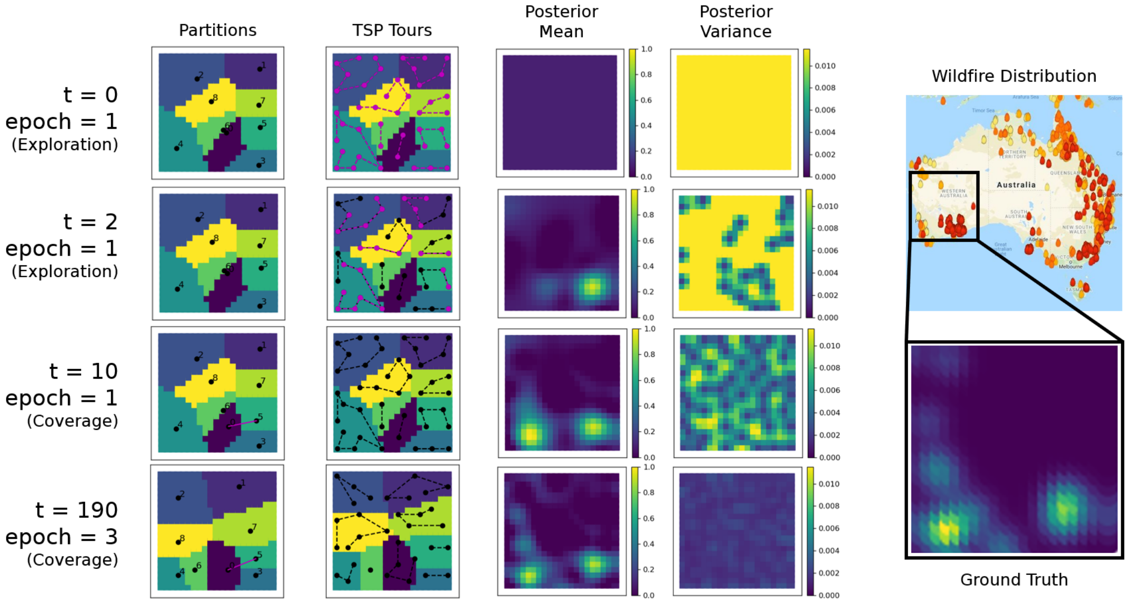


Figure 7.1: Distributed implementation of DSLC.

In each simulation, nine agents are placed uniformly at random over the grid and execute 3 epochs of length 16, 46, and 128 to achieve adaptive coverage of the environment. Partitions are initialized by iterating over the grid and assigning each point to the nearest agent. During the exploration phase of each epoch, partitions are fixed; during the exploitation phase of each epoch, partitions are updated according to the protocol established in [127], where pairwise gossip-based repartitioning occurs between randomly selected neighbors. Coverage cost, regret and maximum variance are computed throughout using (7.2), Definition 7.2, and the maximum diagonal entry of $\mathbf{\Lambda}^{-1}(t)$ from (7.1), respectively. From left to right in turn, Figure 7.1 shows (i) agent positions $\boldsymbol{\eta}_t$ and partitions P_t (ii) TSP sampling tours (iii) posterior mean (iv) variance of $\hat{\phi}$. Pairwise partition updates between gossiping agents are denoted by magenta lines in the leftmost column of panels. Points along TSP tours in the second-from-leftmost column of panels are plotted in magenta prior

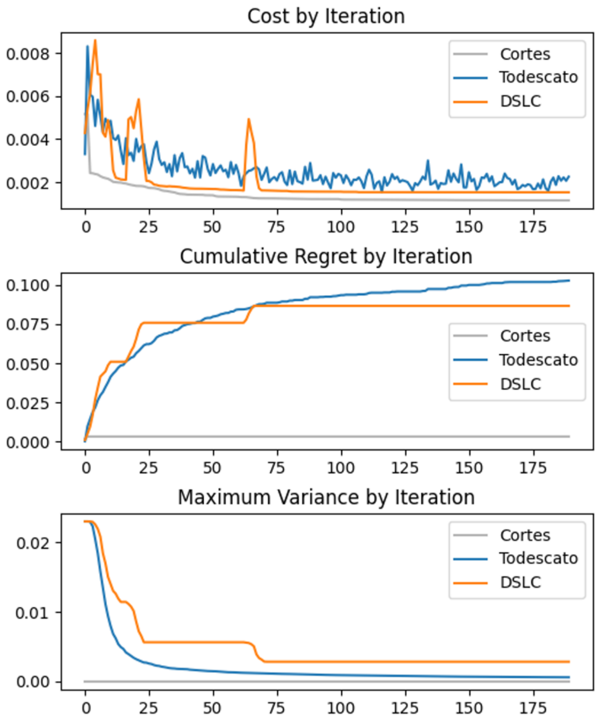


Figure 7.2: Comparison of DSLC, `Todescato` and `Cortes`.

to sampling, and in black after sampling. A simulation video is available online.[2]

The demand function ϕ is normalized in the range $[0, 1]$ and sampled by agents with Gaussian noise parameterized by mean and standard deviation $\mu = 0$, $\sigma = 0.1$. A global Gaussian Process model is assumed to simplify the estimation of $\hat{\phi}$ throughout the simulation, though in practice estimation of $\hat{\phi}$ could be implemented in a fully distributed manner by assuming each agent maintains its own model of $\hat{\phi}$ and employing an information propagation phase described in Section 7.2. Setting the parameter $\alpha = 0.5$ to reduce uncertainty by half within each epoch, $\beta = \alpha^{-3/2}$ is fully determined by Theorem 7.4. Figure 7.1 visualizes the simulation of DSLC.

Figure 7.2 compares the evolution of the coverage regret, coverage cost, and uncertainty reduction in DSLC with that of algorithms proposed in [13] and [68], denoted `Cortes` and `Todescato`, respectively. Agents in [13] are assumed to have perfect knowledge of ϕ and simply go to the centroid of their cell at each iteration; in [68], agents follow a stochastic sampling approach with the probability of exploration proportional to posterior variance in the estimate $\hat{\phi}$ at each iteration. All

[2] `https://youtu.be/nalwrZC6GiI`

results are averaged over 16 simulations of 190 iterations, aligned with the three-implementation of DSLC with epoch lengths 16, 46, and 128. It can be noticed that DSLC empirically achieves sublinear regret. Spikes in regret occur during the exploration phase of each epoch, before agents converge to a pairwise-optimal coverage configuration with respect to $\hat{\phi}$ during the exploitation phase.

Though we do not include the algorithm in our simulations, it is worth noting that DSLC operates in a manner similar to that proposed in [64] where agents spend a number of iterations sampling ϕ to reduce maximum posterior variance $\max_{i\in\{1,\dots,|V|\}} \sigma_i^2(n)$ below a prespecified threshold, then transition to perform coverage for all remaining iterations. Indeed, this algorithm is essentially a special case of DSLC with one epoch and can therefore be expected to perform similarly from an empirical perspective.

7.5 Summary

We propose an adaptive coverage algorithm DSLC that balances the exploration versus exploitation trade-off in learning ϕ and achieving environmental coverage. Our algorithm schedules learning and coverage epochs such that its emphasis gradually shifts from exploration to exploitation while never fully ceasing to learn. Most importantly, we introduce a novel coverage regret that characterizes the deviation of agent configurations and partitions from a centroidal Voronoi partition and derive analytic bounds on the expected cumulative regret for DSLC. In particular, we prove that DSLC will achieve sublinear expected cumulative regret under minor assumptions. The efficacy of DSLC is illustrated through extensive simulation and comparison with existing state-of-the-art approaches to adaptive coverage.

7.6 Bibliographic Remarks

Classical approaches to coverage control [13, 61–63] assume a priori knowledge of ϕ and employ Lloyd's algorithm [132] to drive agents to a local minimum of the coverage cost. In these algorithms, each agent communicates with the agents in the neighboring partitions at each time and updates its partition. Distributed gossip-based coverage algorithms [133] address potential communication bottlenecks in classical approaches by updating partitions pairwise between the agents in neighboring partitions.

While much of the work in coverage considers continuous convex environments, the global convergence property remains an open problem. The asymptotic convergence to a local minimum is normally based on an unproven assumption that there exist finite local optimal points [13, 134]. A discrete graph representation of the environment is considered in [127], which not only enables finite time convergence but also allows for non-convex environments. As has been mentioned earlier in this chapter, the gossip-based coverage algorithm in the graph environment has been proved to converge almost surely to pairwise-optimal partitions in finite time [127].

Recent works have put more focus on the problem of adaptive coverage, in which agents are not assumed to know ϕ a priori. Parametric estimation approaches to adaptive coverage [135, 136] model ϕ as a linear combination of some basis functions and propose algorithms to learn the weights of basis functions; while non-parametric approaches [64–69] model ϕ as the realization of a Gaussian Process and make predictions by conditioning on observed values of ϕ sampled over the operating environment. Alternative approaches to adaptive coverage [137, 138] have also been considered.

A non-parametric adaptive coverage algorithm with provable regret guarantees was presented in this chapter. Similar adaptive coverage algorithms with formal performance guarantees are also developed in [68, 69]. Todescato *et al.* [68] use a Bernoulli random variable for each robot to decide between learning and coverage steps. The distribution of this random variable is designed to ensure the convergence of the algorithm. In contrast, we leverage the so-called “doubling trick" from the MAB literature to design a deterministic schedule of learning and coverage. This allows us to derive formal regret bounds on our adaptive coverage algorithm.

The most closely related work to the ideas presented in this chapter is by Benevento *et al.* [69], which uses a Gaussian process optimization-based [44] approach to design an adaptive coverage algorithm. They derived an upper bound on a notion of coverage regret different from Definition 7.2

in this chapter. Their result is based on a strong assumption that the coverage control algorithm can drive the system to the global minimum of the coverage cost. In contrast, our coverage regret is defined with respect to the local minima which can be achieved by many state-of-the-art coverage control policies including the classic Lloyd's algorithm. By analyzing the coverage regret defined in this chapter, the convergence of the adaptive coverage control can still be shown without requiring the global optimal assumption.

CHAPTER 8

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation has focused upon optimal decision-making in the face of uncertainty. In particular, we address the exploration versus exploitation dilemma in the MAB setup as well as in robotic problems including target search and adaptive coverage control. All proposed algorithms are accompanied by rigorous analysis to indicate their convergence properties.

Since the MAB problem provides a concise mathematical formulation of the exploration versus exploitation dilemma, we have investigated a variety of MAB problem variations that capture different properties of the stochastic environment in real-world problems. For the heavy-tailed bandit, we proposed the Robust MOSS algorithm, which is the first to achieve order optimal worst-case regret while maintaining a logarithm asymptotic regret. For the nonstationary bandits, we studied both the piece-wise stationary bandit and the more general nonstationary bandits with a variation budget. Exact order optimal or near order optimal algorithms for these problem setups are proposed, analyzed, and compared extensively in simulations. As an extension of the single-player MAB, we studied the distributed multi-player bandit in a piece-wise stationary environment. By modifying the single-player policy, novel multi-player policies are designed and proved to maintain group regrets matching with the standard single-player regret even without communication between agents.

For the robotic target search problem, we have considered a scenario in with a robot operates in a 3D environment to search targets on a 2D floor. The target search task is modeled as a hot-spots identification problem in which sensing information is compromised by measurement noise. Since sensing at a location farther from the floor provides better coverage of the environment but less accurate results, we have modeled the sensing field with a multi-fidelity Gaussian process that captures the coverage-accuracy trade-off. Leveraging this novel sensing model, we established a new informative path planning strategy that allows for jointly planning for sampling locations and associated fidelity levels, and thus reduces target search time.

For the adaptive coverage problem, the demand for robotic service within the environment is modeled as a realization of the Gaussian process. With Bayesian techniques, we have devised a policy that balances the tradeoff between learning the demand and covering the environment. To provide analytical rigor, we have defined the coverage regret, and based on it, we have analyzed the convergence property of the proposed online estimation and coverage algorithm.

There are several possible avenues of future studies on problems addressed in this dissertation. The distributed multi-player MAB problem in a general nonstationary environment is a challenging problem and is expected to be applied in the opportunistic spectrum access wherein the stochastic nature of the channel vacancy changes with time. We intend to design a multi-player policy that actively detects the drift in stochastic reward processes such that the players require no prior information about the nonstationary environment. It can be foreseen that the nonstationarity could bring more difficulty in reward estimation so that achieving coordinate behavior to reduce collisions is a trickier task. To deal with it, we can allow communication among agents. For example, the agents can do cooperative reward estimation through a bi-directional communication network by running consensus algorithms as in [27]. Since communicating sampling results may require relatively large bandwidth, to reduce the communication requirement, it is also possible to only require the player to share the ranking of arms as mentioned in [139]. Other issues for such a problem include privacy and defense against adversarial attacks.

We are also interested in extending the single-robot target search policy to cooperative multi-robot search scenarios. As has been mentioned, coverage control is a potential method that can balance the workload. Note that the workload distribution in the environment changes as the search mission progresses, so the robots need to cover the dynamic demands of service. For such a problem, providing analytical rigor would be of interest. Other workload-balancing ideas include using multi-robot path planning methods that solve a vehicle routing problem [140] or orienteering problem [141]. Also of interest would be the implementation of target search algorithms in underwater multi-target search testbeds.

It would be worthwhile to pursue the adaptive coverage problem from a variety of new directions.

The proposed online estimation and coverage algorithm requires an information propagation phase to maintain uniform estimation of demands among agents, while we envision a fully distributed policy that allows for small differences in demand estimates. Besides, the problem setup can be generalized by considering heterogeneous agents that can provide multiple types of services. The quality of service could depend on both the servicing agent and the service type. Considering both inter-service dependencies and the spatial correlations, the multi-task Gaussian process might be a good fit to model demands of different service types. Another interesting direction could be to use the time-varying Gaussian process to model a dynamic environment in which the demands change with time.

Like adaptive coverage control, adaptive multi-robot patrolling is also an interesting problem with the explore-exploit tradeoff. In the multi-robot patrolling problem [142], a team of robots circling around a set of important locations (viewpoints) with different known priorities. The objective is to minimize the weighted refresh time, which is the longest time interval between any two visits of a viewpoint, weighted by the viewpoint's priority. We envision addressing this problem by considering the priorities of viewpoints to be unknown and time-varying so that they need to be learned. Since the LM-DSEE algorithm for the piece-wise stationary MAB problem has a block allocation structure that benefits path planning, its multi-player extension and associated distributed control methods could be promising to solve the problem.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[2] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, p. 47, 2002.

[3] M. Vidyasagar, "Law of large numbers, heavy-tailed distributions, and the recent financial crisis," in *Perspectives in Mathematical System Theory, Control, and Signal Processing*. Springer, 2010, pp. 285–295.

[4] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[5] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[6] O. Avner and S. Mannor, "Concurrent bandits and cognitive radio networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 66–81.

[7] J. R. Krebs, A. Kacelnik, and P. Taylor, "Test of optimal sampling by foraging great tits," *Nature*, vol. 275, no. 5675, pp. 27–31, 1978.

[8] V. Srivastava, P. Reverdy, and N. E. Leonard, "On optimal foraging and multi-armed bandits," in *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, 2013, pp. 494–499.

[9] V. Srivastava, P. Reverdy, and N. E. Leonard, "Surveillance in an abruptly changing world via multiarmed bandits," in *IEEE Conference on Decision and Control*, 2014, pp. 692–697.

[10] M. Y. Cheung, J. Leighton, and F. S. Hover, "Autonomous mobile acoustic relay positioning as a multi-armed bandit with switching costs," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, Nov. 2013, pp. 3368–3373.

[11] Y. Sung, D. Dixit, and P. Tokekar, "Environmental hotspot identification in limited time with a uav equipped with a downward-facing camera," *arXiv preprint arXiv:1909.08483*, 2019.

[12] M. C. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, vol. 87, no. 1, pp. 1–13, 2000.

[13] J. Cortés, S. Martínez, T. Karataş, and F. Bullo, “Coverage control for mobile sensing networks,” *IEEE Transactions on Robotics and Automation*, vol. 20, no. 2, pp. 243–255, 2004.

[14] H. Robbins, “Some aspects of the sequential design of experiments,” *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[15] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[16] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[17] A. Garivier and O. Cappé, “The KL-UCB algorithm for bounded stochastic bandits and beyond,” in *Proceedings of the 24th Conference on Learning Theory*, vol. 19, Budapest, Hungary, 2011, pp. 359–376.

[18] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, “Bandits with heavy tail,” *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7711–7717, 2013.

[19] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “The nonstochastic multiarmed bandit problem,” *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[20] A. Garivier and E. Moulines, “On upper-confidence bound policies for switching bandit problems,” in *International Conference on Algorithmic Learning Theory*. Springer, 2011, pp. 174–188.

[21] O. Besbes and Y. Gur, “Stochastic multi-armed-bandit problem with non-stationary rewards,” in *Advances in neural information processing systems*, 2014, pp. 199–207.

[22] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: IID rewards,” *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.

[23] Y. Gai and B. Krishnamachari, “Distributed stochastic online learning policies for opportunistic spectrum access,” *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6184–6193, 2014.

[24] D. Kalathil, N. Nayyar, and R. Jain, “Decentralized learning for multiplayer multiarmed bandits,” *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[25] N. Nayyar, D. Kalathil, and R. Jain, “On regret-optimal learning in decentralized multiplayer multiarmed bandits,” *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 597–606, 2018.

[26] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, “Multi-armed bandits in multi-agent networks,” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[27] P. Landgren, V. Srivastava, and N. E. Leonard, “On distributed cooperative decision-making in multiarmed bandits,” in *2016 European Control Conference*, Aalborg, Denmark, 2016, pp. 243–248.

[28] P. Landgren, V. Srivastava, and N. E. Leonard, “Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms,” in *IEEE Conference on Decision and Control*, Las Vegas, NV, USA, Dec. 2016, pp. 167–172.

[29] A. B. H. Alaya-Feki, E. Moulines, and A. LeCornec, “Dynamic spectrum access with non-stationary multi-armed bandit,” in *IEEE Workshop on Signal Processing Advances in Wireless Communications*, 2008, pp. 416–420.

[30] Y. Li, Q. Hu, and N. Li, “A reliability-aware multi-armed bandit approach to learn and select users in demand response,” *Automatica*, vol. 119, p. 109015, 2020.

[31] D. Kalathil and R. Rajagopal, “Online learning for demand response,” in *Annual Allerton Conference on Communication, Control, and Computing*, 2015, pp. 218–222.

[32] D. Agarwal, B. C. Chen, P. Elango, N. Motgi, S. T. Park, R. Ramakrishnan, S. Roy, and J. Zachariah, “Online models for content optimization,” in *Advances in Neural Information Processing Systems*, vol. 21. Curran Associates, Inc., 2009, pp. 17–24.

[33] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A contextual-bandit approach to personalized news article recommendation,” in *International Conference on World Wide Web*, 2010, pp. 661–670.

[34] C. Baykal, G. Rosman, S. Claici, and D. Rus, “Persistent surveillance of events with unknown, time-varying statistics,” in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2017, pp. 2682–2689.

[35] R. Dimitrova, I. Gavran, R. Majumdar, V. S. Prabhu, and S. E. Z. Soudjani, “The robot routing problem for collecting aggregate stochastic rewards,” *arXiv preprint arXiv:1704.05303*, 2017.

[36] V. Srivastava, F. Pasqualetti, and F. Bullo, “Stochastic surveillance strategies for spatial quickest detection,” *The International Journal of Robotics Research*, vol. 32, no. 12, pp. 1438–1458, 2013.

[37] R. Agrawal, M. V. Hedge, and D. Teneketzis, “Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost,” *IEEE Transactions on Automatic Control*, vol. 33, no. 10, pp. 899–906, 1988.

[38] P. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized Gaussian multiarmed bandits," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 544–571, 2014.

[39] V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg, "Batched bandit problems," *The Annals of Statistics*, vol. 44, no. 2, pp. 660–681, 2016.

[40] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, 2013.

[41] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

[42] C. K. Williams and C. E. Rasmussen, *Gaussian processes for Machine Learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[43] S. Vasudevan, F. Ramos, E. Nettleton, and H. Durrant-Whyte, "Gaussian process modeling of large-scale terrain," *Journal of Field Robotics*, vol. 26, no. 10, pp. 812–840, 2009.

[44] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[45] G. A. Hollinger, B. Englot, F. S. Hover, U. Mitra, and G. S. Sukhatme, "Active planning for underwater inspection and the benefit of adaptivity," *The International Journal of Robotics Research*, vol. 32, no. 1, pp. 3–18, 2013.

[46] D. E. Soltero, M. Schwager, and D. Rus, "Generating informative paths for persistent sensing in unknown environments," in *IEEE/RSJ Int Conf on Intelligent Robots and Systems*, Vilamoura, Algarve, Portugal, Oct. 2012, pp. 2172–2179.

[47] J. Yu, M. Schwager, and D. Rus, "Correlated orienteering problem and its application to informative path planning for persistent monitoring tasks," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 342–349.

[48] G. A. Hollinger and G. S. Sukhatme, "Sampling-based robotic information gathering algorithms," *The International Journal of Robotics Research*, vol. 33, no. 9, pp. 1271–1287, 2014.

[49] G. Hitz, E. Galceran, M.-È. Garneau, F. Pomerleau, and R. Siegwart, "Adaptive continuous-space informative path planning for online environmental monitoring," *Journal of Field Robotics*, vol. 34, no. 8, pp. 1427–1449, 2017.

[50] G. Hitz, A. Gotovos, M.-É. Garneau, C. Pradalier, A. Krause, R. Y. Siegwart et al., “Fully autonomous focused exploration for robotic environmental monitoring,” in *IEEE International Conference on Robotics and Automation*, 2014, pp. 2658–2664.

[51] N. Atanasov, J. Le Ny, K. Daniilidis, and G. J. Pappas, “Information acquisition with sensing robots: Algorithms and error bounds,” in *IEEE International Conference on Robotics and Automation*, 2014, pp. 6447–6454.

[52] A. A. Meera, M. Popović, A. Millane, and R. Siegwart, “Obstacle-aware adaptive informative path planning for uav-based target search,” in *IEEE International Conference on Robotics and Automation*, 2019, pp. 718–724.

[53] X. Lan and M. Schwager, “Planning periodic persistent monitoring trajectories for sensing robots in Gaussian random fields,” in *IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, May 2013, pp. 2415–2420.

[54] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. E. Davis, “Collective motion, sensor networks, and ocean sampling,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 48–74, 2007.

[55] S. L. Smith, M. Schwager, and D. Rus, “Persistent robotic tasks: Monitoring and sweeping in changing environments,” *IEEE Transactions on Robotics*, vol. 28, no. 2, pp. 410–426, 2012.

[56] C. G. Cassandras, X. Lin, and X. Ding, “An optimal control approach to the multi-agent persistent monitoring problem,” *IEEE Transactions on Automatic Control*, vol. 58, no. 4, pp. 947–961, 2013.

[57] R. N. Smith, M. Schwager, S. L. Smith, B. H. Jones, D. Rus, and G. S. Sukhatme, “Persistent ocean monitoring with underwater gliders: Adapting sampling resolution,” *Journal of Field Robotics*, vol. 28, no. 5, pp. 714–741, 2011.

[58] A. Krause and C. E. Guestrin, “Near-optimal nonmyopic value of information in graphical models,” in *Proceedings of the 21st Conference Conference on Uncertainty in Artificial Intelligence*, Edinburgh, Scotland, July 2005, pp. 324–331.

[59] P. Auer and R. Ortner, “UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem,” *Periodica Mathematica Hungarica*, vol. 61, no. 1-2, pp. 55–65, 2010.

[60] S. Kalyanakrishnan and P. Stone, “Efficient selection of multiple bandit arms: Theory and practice,” in *ICML*, 2010.

[61] J. Cortés and F. Bullo, “Coordination and Geometric Optimization via Distributed Dynamical Systems,” *SIAM Journal on Control and Optimization*, vol. 44, no. 5, pp. 1543–1574, 2005.

[62] F. Lekien and N. E. Leonard, "Nonuniform coverage and cartograms," *IEEE Conference on Decision and Control*, pp. 5518–5523, 2010.

[63] I. I. Hussein and D. M. Stipanovic, "Effective coverage control for mobile sensor networks with guaranteed collision avoidance," *IEEE Transactions on Control Systems Technology*, vol. 15, no. 4, pp. 642–657, 2007.

[64] J. Choi, J. Lee, and S. Oh, "Swarm intelligence for achieving the global maximum using spatio-temporal Gaussian processes," *Proceedings of the American Control Conference*, pp. 135–140, 2008.

[65] Y. Xu and J. Choi, "Adaptive sampling for learning Gaussian processes using mobile sensor networks," *Sensors*, vol. 11, no. 3, pp. 3051–3066, 2011.

[66] W. Luo and K. Sycara, "Adaptive Sampling and Online Learning in Multi-Robot Sensor Coverage with Mixture of Gaussian Processes," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 6359–6364.

[67] W. Luo, C. Nam, G. Kantor, and K. Sycara, "Distributed environmental modeling and adaptive sampling for multi-robot sensor coverage," in *International Joint Conference on Autonomous Agents and Multiagent Systems*, 2019, pp. 1488–1496.

[68] M. Todescato, A. Carron, R. Carli, G. Pillonetto, and L. Schenato, "Multi-robots Gaussian estimation and coverage control: From client–server to peer-to-peer architectures," *Automatica*, vol. 80, pp. 284–294, 2017.

[69] A. Benevento, M. Santos, G. Notarstefano, K. Paynabar, M. Bloch, and M. Egerstedt, "Multi-robot coordination for estimation and coverage of unknown spatial fields," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 7740–7746.

[70] J. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proceedings of the 22nd conference on learning theory*, 2009, pp. 217–226.

[71] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for sequential allocation problems," *Advances in Applied Mathematics*, vol. 17, no. 2, pp. 122–142, 1996.

[72] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *IEEE Annual Foundations of Computer Science*, 1995, pp. 322–331.

[73] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *Journal of Machine Learning Research*, vol. 5, no. Jun, pp. 623–648, 2004.

[74] L. Wei and V. Srivastava, "Minimax policy for heavy-tailed bandits," *IEEE Control Systems Letters*, vol. 5, no. 4, pp. 1423–1428, 2021.

[75] X. Fan, I. Grama, and Q. Liu, “Hoeffding’s inequality for supermartingales,” *Stochastic Processes and their Applications*, vol. 122, no. 10, pp. 3545–3559, 2012.

[76] S. Bubeck, “Bandits games and clustering foundations,” Ph.D. dissertation, Université des Sciences et Technologie de Lille - Lille I, 2010.

[77] O. C. E. Kaufmann and A. Garivier, “On bayesian upper confidence bounds for bandit problems,” in *Artificial Intelligence and Statistics*, 2012, pp. 592–600.

[78] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in *Conference on Learning Theory*, 2012, pp. 39–1.

[79] N. K. E. Kaufmann and R. Munos, “Thompson sampling: An asymptotically optimal finite-time analysis,” in *International Conference on Algorithmic Learning Theory*. Springer, 2012, pp. 199–213.

[80] P. Ménard and A. Garivier, “A minimax and asymptotically optimal algorithm for stochastic bandits,” in *Algorithmic Learning Theory*, vol. 76, 2017.

[81] R. Degenne and V. Perchet, “Anytime optimal algorithms in stochastic multi-armed bandits,” in *International Conference on Machine Learning*, 2016, pp. 1587–1595.

[82] L. Wei and V. Srivastava, “On abruptly-changing and slowly-varying multiarmed bandit problems,” in *American Control Conference*, Milwaukee, WI, June 2018, pp. 6291–6296.

[83] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[84] J. C. Gittins, “Bandit processes and dynamic allocation indices,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 2, pp. 148–164, 1979.

[85] K. Liu and Q. Zhao, “Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access,” *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[86] L. Kocsis and C. Szepesvári, “Discounted UCB,” in *2nd PASCAL Challenges Workshop*, vol. 2, 2006.

[87] C. Hartland, N. Baskiotis, S. Gelly, M. Sebag, and O. Teytaud, “Change Point Detection and Meta-Bandits for Online Learning in Dynamic Environments,” in *Conférence Francophone Sur L’Apprentissage Automatique*, Grenoble, France, July 2007, pp. 237–250.

[88] F. Liu, J. Lee, and N. Shroff, “A change-detection based framework for piecewise-stationary multi-armed bandit problem,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[89] L. Besson and E. Kaufmann, “The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits,” *arXiv preprint arXiv:1902.01575*, 2019.

[90] Y. Cao, Z. Wen, B. Kveton, and Y. Xie, “Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit,” in *International Conference on Artificial Intelligence and Statistics*, 2019, pp. 418–427.

[91] R. Allesiardo and R. Féraud, “Exp3 with drift detection for the switching bandit problem,” in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2015, pp. 1–7.

[92] J. Mellor and J. Shapiro, “Thompson sampling in switching environments with bayesian online change detection,” in *Artificial Intelligence and Statistics*, 2013, pp. 442–450.

[93] P. Auer, P. Gajane, and R. Ortner, “Adaptively tracking the best bandit arm with an unknown number of distribution changes,” in *Annual Conference on Learning Theory*, 2019, pp. 138–158.

[94] Y. Chen, C. W. Lee, H. Luo, and C. Y. Wei, “A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free,” in *Proceedings of the 32nd Conference on Learning Theory*, vol. 99, Phoenix, USA, 2019, pp. 696–726.

[95] L. Wei and V. Srivastava, “On distributed multi-player multiarmed bandit problems in abruptly changing environment,” in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 5783–5788.

[96] R. Burkard, M. Dell’Amico, and S. Martello, *Assignment problems: revised reprint*. SIAM, 2012.

[97] D. P. Bertsekas, “The auction algorithm: A distributed relaxation method for the assignment problem,” *Annals of operations research*, vol. 14, no. 1, pp. 105–123, 1988.

[98] E. Boursier and V. Perchet, “SIC-MMAB: Synchronisation involves communication in multiplayer multi-armed bandits,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 2249–2257.

[99] C. Shi and C. Shen, “On no-sensing adversarial multi-player multi-armed bandits with collision communications,” *IEEE Journal on Selected Areas in Information Theory*, 2021.

[100] I. Bistritz and A. Leshem, “Distributed multi-player bandits-a game of thrones approach,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 7222–7232.

[101] J. R. Marden, H. P. Young, and L. Y. Pao, “Achieving pareto optimality through distributed learning,” *SIAM Journal on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, 2014.

[102] O. Besbes, Y. Gur, and A. Zeevi, "Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards," *Stochastic Systems*, vol. 9, no. 4, pp. 319–337, 2019.

[103] V. Raj and S. Kalyani, "Taming non-stationary bandits: A Bayesian approach," *arXiv preprint arXiv:1707.09727*, 2017.

[104] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Learning to optimize under non-stationarity," in *Proceedings of Machine Learning Research*, vol. 89, 16–18 Apr 2019, pp. 1079–1087.

[105] P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou, "A simple approach for non-stationary linear bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 746–755.

[106] Y. Russac, C. Vernade, and O. Cappé, "Weighted linear bandits for non-stationary environments," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 12 017–12 026.

[107] L. Wei, X. Tan, and V. Srivastava, "Expedited multi-target search with guaranteed performance via multi-fidelity Gaussian processes," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, NV (Virtual), Oct. 2020, pp. 7095–7100.

[108] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[109] P. Perdikaris, "Gaussian processes a hands-on tutorial," 2017. [Online]. Available: https://github.com/paraklas/GPTutorial

[110] S. Kemna, J. G. Rogers, C. Nieto-Granda, S. Young, and G. S. Sukhatme, "Multi-robot coordination through dynamic Voronoi partitioning for informative adaptive sampling in communication-constrained environments," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 2124–2130.

[111] D. Applegate, R. Bixby, V. Chvatal, and W. Cook, "Concorde TSP solver," 2006.

[112] J. Y. Audibert and S. Bubeck, "Best arm identification in multi-armed bandits," in *Proceedings of the 23rd Conference on Learning Theory*, 2010, pp. 41–53.

[113] E. Rolf, D. Fridovich-Keil, M. Simchowitz, B. Recht, and C. Tomlin, "A successive-elimination approach to adaptive robotic sensing," *ArXiv e-prints*, 2018.

[114] M. M. M. Manhães, S. A. Scherer, M. Voss, L. R. Douat, and T. Rauschenbach, "UUV simulator: A Gazebo-based package for underwater intervention and multi-robot simulation," in *OCEANS 2016 MTS/IEEE Monterey*. IEEE, 2016, pp. 1–8.

[115] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1964.

[116] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, “An analysis of approximations for maximizing submodular set functions,” *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.

[117] K. Kandasamy, G. Dasarathy, J. B. Oliva, J. Schneider, and B. Póczos, “Gaussian process bandit optimisation with multi-fidelity evaluations,” in *Advances in Neural Information Processing Systems*, 2016, pp. 992–1000.

[118] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[119] H. J. Karloff, “How long can a euclidean traveling salesman tour be?” *SIAM Journal on Discrete Mathematics*, vol. 2, no. 1, pp. 91–99, 1989.

[120] A. Singh, A. Krause, C. Guestrin, and W. J. Kaiser, “Efficient informative sensing using multiple robots,” *Journal of Artificial Intelligence Research*, vol. 34, no. 2, p. 707, 2009.

[121] A. Krause, A. Singh, and C. Guestrin, “Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies,” *Journal of Machine Learning Research*, vol. 9, no. Feb, pp. 235–284, 2008.

[122] J. L. Ny and G. J. Pappas, “On trajectory optimization for active sensing in Gaussian process models,” in *IEEE Conf on Decision and Control and Chinese Control Conference*, Shanghai, China, Dec. 2009, pp. 6286–6292.

[123] S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen, “Combinatorial pure exploration of multi-armed bandits,” in *Advances in Neural Information Processing Systems*, 2014, pp. 379–387.

[124] P. Reverdy, V. Srivastava, and N. E. Leonard, “Satisficing in multi-armed bandit problems,” *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3788 – 3803, 2017.

[125] L. Wei, A. McDonald, and V. Srivastava, “Multi-robot Gaussian process estimation and coverage: a deterministic sequencing algorithm and regret analysis,” in *Proceedings of the IEEE International Conference on Robotics and Automation*, Xi’an, CN (Virtual), 202.

[126] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I : Estimation Theory*. Prentice Hall, 1993.

[127] J. W. Durham, R. Carli, P. Frasca, and F. Bullo, “Discrete partitioning and coverage control for gossiping robots,” *IEEE Transactions on Robotics*, vol. 28, no. 2, pp. 364–378, 2012.

[128] F. Bullo, J. Cortés, and S. Martínez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series. Princeton University Press, 2009, electronically available at http://coordinationbook.info.

[129] H. Lim and C. Kim, "Flooding in wireless ad hoc networks," *Computer Communications*, vol. 24, no. 3-4, pp. 353–363, 2001.

[130] L. Wang and F. Xiao, "Finite-time consensus problems for networks of dynamic agents," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 950–955, 2010.

[131] C. Paradis, "Fires from Space: Australia," 2019. [Online]. Available: https://www.kaggle.com/carlosparadis/fires-from-space-australia-and-new-zeland

[132] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[133] F. Bullo, R. Carli, and P. Frasca, "Gossip coverage control for robotic networks: Dynamical systems on the space of partitions," *SIAM Journal on Control and Optimization*, vol. 50, no. 1, pp. 419–447, 2012.

[134] Q. Du, M. Emelianenko, and L. Ju, "Convergence of the lloyd algorithm for computing centroidal voronoi tessellations," *SIAM journal on numerical analysis*, vol. 44, no. 1, pp. 102–119, 2006.

[135] M. Schwager, D. Rus, and J. J. Slotine, "Decentralized, adaptive coverage control for networked robots," *International Journal of Robotics Research*, vol. 28, no. 3, pp. 357–375, 2009.

[136] M. Schwager, M. P. Vitus, S. Powers, D. Rus, and C. J. Tomlin, "Robust adaptive coverage control for robotic sensor networks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 3, pp. 462–476, 2017.

[137] P. Davison, N. E. Leonard, A. Olshevsky, and M. Schwemmer, "Nonuniform Line Coverage from Noisy Scalar Measurements," *IEEE Transactions on Automatic Control*, vol. 60, no. 7, pp. 1975–1980, 2015.

[138] J. Choi and R. Horowitz, "Learning coverage control of mobile sensing agents in one-dimensional stochastic environments," *IEEE Transactions on Automatic Control*, vol. 55, no. 3, pp. 804–809, 2010.

[139] M. Agarwal, V. Aggarwal, and K. Azizzadenesheli, "Multi-agent multi-armed bandits with limited communication," *arXiv preprint arXiv:2102.08462*, 2021.

[140] P. Toth and D. Vigo, *The vehicle routing problem*. SIAM, 2002.

[141] A. Gunawan, H. C. Lau, and P. Vansteenwegen, “Orienteering problem: A survey of recent variants, solution approaches and applications,” *European Journal of Operational Research*, vol. 255, no. 2, pp. 315–332, 2016.

[142] F. Pasqualetti, J. W. Durham, and F. Bullo, “Cooperative patrolling via weighted tours: Performance analysis and distributed algorithms,” *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1181–1188, 2012.