MORGAN & CLAYPOOL PUBLISHERS

Multi-Armed Bandits

Theory and Applications to Online Learning in Networks

Qing Zhao

Synthesis Lectures on Communication Networks

R. Srikant, *Series Editor*

Multi-Armed Bandits

Theory and Applications to Online Learning in Networks

Synthesis Lectures on Communication Networks

Synthesis Lectures on Communication Networks is an ongoing series of 75- to 150-page publications on topics on the design, implementation, and management of communication networks. Each lecture is a self-contained presentation of one topic by a leading expert. The topics range from algorithms to hardware implementations and cover a broad spectrum of issues from security to multiple-access protocols. The series addresses technologies from sensor networks to reconfigurable optical networks.
The series is designed to:

- Provide the best available presentations of important aspects of communication networks.
- Help engineers and advanced students keep up with recent developments in a rapidly evolving technology.
- Facilitate the development of courses in this field

Multi-Armed Bandits: Theory and Applications to Online Learning in Networks
Qing Zhao
2019

Diffusion Source Localization in Large Networks
Lei Ying and Kai Zhu
2018

Communications Networks: A Concise Introduction, Second Edition
Jean Walrand and Shyam Parekh
2017

BATS Codes: Theory and Practice
Shenghao Yang and Raymond W. Yeung
2017

Analytical Methods for Network Congestion Control
Steven H. Low
2017

Advances in Multi-Channel Resource Allocation: Throughput, Delay, and Complexity
Bo Ji, Xiaojun Lin, and Ness B. Shroff
2016

A Primer on Physical-Layer Network Coding
Soung Chang Liew, Lu Lu, and Shengli Zhang
2015

Sharing Network Resources
Abhay Parekh and Jean Walrand
2014

Wireless Network Pricing
Jianwei Huang and Lin Gao
2013

Performance Modeling, Stochastic Networks, and Statistical Multiplexing, Second Edition
Ravi R. Mazumdar
2013

Packets with Deadlines: A Framework for Real-Time Wireless Networks
I-Hong Hou and P.R. Kumar
2013

Energy-Efficient Scheduling under Delay Constraints for Wireless Networks
Randall Berry, Eytan Modiano, and Murtaza Zafer
2012

NS Simulator for Beginners
Eitan Altman and Tania Jiménez
2012

Network Games: Theory, Models, and Dynamics
Ishai Menache and Asuman Ozdaglar
2011

An Introduction to Models of Online Peer-to-Peer Social Networking
George Kesidis
2010

Stochastic Network Optimization with Application to Communication and Queueing Systems
Michael J. Neely
2010

Scheduling and Congestion Control for Wireless and Processing Networks
Libin Jiang and Jean Walrand
2010

Performance Modeling of Communication Networks with Markov Chains
Jeonghoon Mo
2010

Communication Networks: A Concise Introduction
Jean Walrand and Shyam Parekh
2010

Path Problems in Networks
John S. Baras and George Theodorakopoulos
2010

Performance Modeling, Loss Networks, and Statistical Multiplexing
Ravi R. Mazumdar
2009

Network Simulation
Richard M. Fujimoto, Kalyan S. Perumalla, and George F. Riley
2006

Multi-Armed Bandits: Theory and Applications to Online Learning in Networks

Qing Zhao

www.morganclaypool.com

ISBN: 9781627056380 paperback
ISBN: 9781627058711 ebook
ISBN: 9781681736372 hardcover

DOI 10.2200/S00941ED2V01Y201907CNT022

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON COMMUNICATION NETWORKS

Lecture #22
Series Editor: R. Srikant, *University of Illinois at Urbana-Champaign*
Founding Editor Emeritus: Jean Walrand, *University of California, Berkeley*
Series ISSN
Print 1935-4185 Electronic 1935-4193

Multi-Armed Bandits

Theory and Applications to Online Learning in Networks

Qing Zhao
Cornell University

SYNTHESIS LECTURES ON COMMUNICATION NETWORKS #22

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Multi-armed bandit problems pertain to optimal sequential decision making and learning in unknown environments. Since the first bandit problem posed by Thompson in 1933 for the application of clinical trials, bandit problems have enjoyed lasting attention from multiple research communities and have found a wide range of applications across diverse domains. This book covers classic results and recent development on both Bayesian and frequentist bandit problems. We start in Chapter 1 with a brief overview on the history of bandit problems, contrasting the two schools—Bayesian and frequentist—of approaches and highlighting foundational results and key applications. Chapters 2 and 4 cover, respectively, the canonical Bayesian and frequentist bandit models. In Chapters 3 and 5, we discuss major variants of the canonical bandit models that lead to new directions, bring in new techniques, and broaden the applications of this classical problem. In Chapter 6, we present several representative application examples in communication networks and social-economic systems, aiming to illuminate the connections between the Bayesian and the frequentist formulations of bandit problems and how structural results pertaining to one may be leveraged to obtain solutions under the other.

KEYWORDS

multi-armed bandit, machine learning, online learning, reinforcement learning, Markov decision processes

To Peter Whittle
and to Lang and Everett.

Contents

Preface

The term "multi-armed bandit" comes from likening an archetypal online learning problem to playing a slot machine that has multiple arms (slot machines are also known as bandits due to their ability to empty the player's pockets). Each arm, when pulled, generates random rewards drawn from an unknown distribution or a known distribution with an unknown mean. The player chooses one arm to pull at each time, with the objective of accumulating, in expectation, as much reward as possible over a given time horizon. The tradeoff facing the player is a classic one, that is, to explore a less observed arm which may hold a greater potential for the future or to exploit an arm with a history of offering good rewards. It is this tension between learning and earning that lends complexity and richness to the bandit problems.

As in many problems involving unknowns, bandit problems can be treated within the Bayesian or frequentist frameworks, depending on whether the unknowns are viewed as random variables with known prior distributions or as deterministic quantities. These two schools have largely evolved independently. In recent years, we witness increased interests and much success in cross-pollination between the two schools. It is my hope that by covering both the Bayesian and frequentist bandit models, this book further stimulates research interests in this direction.

We start in Chapter 1 with an overview on the history and foundational results of the bandit problems within both frameworks. In Chapters 2 and 4, we devote our attention to the canonical Bayesian and frequentist formulations. Major results are treated in detail. Proofs for key theorems are provided.

New and emerging applications in computer science, engineering, and social-economic systems give rise to a diverse set of variants of the classical models, generating new directions and bringing in new techniques to this classical problem. We discuss major variants under the Bayesian framework and the frequentist framework in Chapters 3 and 5, respectively. The coverage, inevitably incomplete, focuses on the general formulations and major results with technical details often omitted. Special attention is given to the unique challenges and additional structures these variants bring to the original bandit models. Being derivative to the original models, these variants also offer a deeper appreciation and understanding of the core theory and techniques. In addition to bringing awareness of new bandit models and providing reference points, these two chapters point out unexplored directions and open questions.

In Chapter 6, we present application examples of the bandit models in communication networks and social-economic systems. While these examples provide only a glimpse of the expansive range of potential applications of bandit models, it is my hope that they illustrate two fruitful research directions: applications with additional structures that admit stronger results than what can be offered by the general theory, and applications bringing in new objectives and constraints that push the boundaries of the bandit models. These examples are chosen also to show the connections between the Bayesian and frequentist formulations and how structural results pertaining to one may be leveraged to obtain solutions under the other.

Qing Zhao
Ithaca, NY, August 2019

Acknowledgments

In December 2015, Srikant, the editor of the series, asked whether I would be interested in writing a book on multi-armed bandits. By that time, I had worked on bandit problems for a decade, starting with the Bayesian and then the frequentist. I was quite confident in taking on the task and excited with the ambition of bringing together the two schools of approaches together within one book, which I felt was lacking in the literature and was much needed. When asked of a timeframe for finishing the book, I gave an estimate of one year. "That ought to leave me plenty of margin." I thought. My son, Everett, was one year old then.

Everett is starting kindergarten next week.

Writing this book has been a humbling experience. The vast landscape of the existing literature, both classical and new, reincarnations of ideas, often decades apart, and quite a few reinventions of wheels (with contributions from myself in that regard), have made the original goal of giving a comprehensive coverage and respecting the historical roots of all results seem unattainable at times. If it were not for the encouragement and persistent nudging from Srikant and the publisher Michael Morgan, the book would have remained unfinished forever. I do not think I have achieved the original goal. This is a version I can, at least, live with.

Many people have helped me in learning this fascinating subject. The first paper I read on bandit problems was "Playing Golf with Two Balls" pointed to me by Vikram Krishnamurthy, then a professor at UBC and now my colleague at Cornell. It was the summer of 2005 when I visited Vikram. We were working on a sensor scheduling problem under the objective of network lifetime maximization, which leads to a stochastic shortest-path bandit. The appeal of the bandit problems was instantaneous and has never faded, and I must admit a stronger affection towards the Bayesian models, likely due to the earlier exposure. Special thanks go to Peter Whittle of the University of Cambridge. I am forever grateful to his tremendous encouragement through my career and generous comments on our results on restless bandits. His incisive writing has always been an inspiration. Many thanks to my students, past and current, who taught me most things I know about bandits through presentations in our endless group meetings and through their research, in particular, Keqin Liu and Sattar Vakili whose dissertations focused almost exclusively on bandit problems.

My deepest appreciation goes to my husband, Lang, for letting me hide away for weeks finishing up a first draft while he took care of Everett, for agreeing to read the draft and providing comments and actually did so for the Introduction! I thank my dear Everett for the many hours sitting patiently next to me, copying on his iPad every letter I typed. It was this July in Sweden when there was no daycare and I was trying to wrap up the book. He has been the most faithful reader of the book, who read, not word by word, but letter by letter.

I thank the U.S. National Science Foundation, the Army Research Office, and the Army Research Lab for supporting the research of my group and the talented Ph.D. students I worked with. I am grateful to the support during my 2018–2019 sabbatical leave in Sweden from the European Union through a Marie Skłodowska-Curie grant[1] and from the Chalmers University of Technology through a Jubilee Professorship. Their generous support has allowed me to put in continuous effort on the book and finally push it over the finish line.

Qing Zhao
Ithaca, NY, August 2019

[1] Current supporting grants include the National Science Foundation Grant CCF-1815559, the Army Research Laboratory Network Science CTA under Cooperative Agreement W911NF-09-2-0053, and the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754412.

CHAPTER 1 Introduction

We start with a brief overview of the multi-armed bandit problems, contrasting two schools of approaches—Bayesian and frequentist—and highlighting foundational results and key applications. Concepts introduced informally here will be treated with care and rigor in later chapters.

1.1 MULTI-ARMED BANDIT PROBLEMS

The first multi-armed bandit problem was posed by Thompson, 1933 [190], for the application of clinical trial. The problem considered there is as follows. Suppose that two experimental treatments are available for a certain disease. The effectiveness of these two treatments is unknown. The decision on which treatment to use on each patient is made sequentially in time based on responses of past patients to their prescribed treatment. In Thompson's study, a patient's response to a treatment is assumed to be dichotomous: success or failure. The two treatments are then equated with two Bernoulli random variables with unknown parameters θ_1 and θ_2. The objective is to prescribe to as many patients as possible the treatment that has a higher mean value.

In earlier studies, this type of sequential learning problems was referred to as "sequential design of experiments" (see, for example, Robbins, 1952 [168] and Bellman, 1956 [30]). The term "multi-armed bandit" first appeared in several papers published in the late 1950s to early 1960s (see Bradt, Johnson, and Karlin, 1956 [43], Vogel, 1960a, 1960b [203, 204], and Feldman, 1962 [79]), by likening the problem to playing a bandit slot machine with multiple arms, each generating random rewards drawn from a distribution with an unknown mean θ_i.

Bandit problems arise in a diverse range of applications. For instance, in communication networks, the problem may be in the form of which channels to access, which paths to route packets, which queues to serve, or which sensors to activate. For applications in social and economic systems, arms may represent products or movies to recommend, prices to set for a new product, or documents and sponsored ads to display on a search engine. The classical problem of stochastic optimization concerned with the minimization of an unknown random loss function can also be viewed as a bandit problem in disguise. In this case, arms, given by the domain of the loss function, are no longer discrete.

1.2 AN ESSENTIAL CONFLICT: EXPLORATION VS. EXPLOITATION

The essence of the multi-armed bandit problems is in the tradeoff between exploitation and exploration where the player faces the conflicting objectives of playing the arm with the best reward history and playing a less explored arm to learn its reward mean so that better decisions can be made in the future. This essential conflict was well articulated by Whittle in the foreword to the first edition of Gittins's monograph on multi-armed bandits (1989 [89]): "Bandit problems embody in essential form a conflict evident in all human action: choosing actions which yield immediate reward vs. choosing actions (e.g., acquiring information or preparing the ground) whose benefit will come only later."

This tradeoff between exploitation and exploration was sharply illustrated with a simple example in the monograph by Berry and Fristedt, 1985 [33]. Suppose there are two coins with bias θ_1 and θ_2. It is known that $\theta_1 = \frac{1}{2}$ (a fair coin) and θ_2 is either 1 (a coin with two heads) or 0 (a coin with two tails) with probability $\frac{1}{4}$ and $\frac{3}{4}$, respectively. At each time, one coin is selected and flipped. The objective is to maximize the expected number of heads in T coin flips. Since the immediate reward from flipping coin 1 is $\frac{1}{2}$ comparing with $\frac{1}{4}$ from flipping coin 2, a myopic strategy that aims solely at maximizing immediate reward would flip coin 1 indefinitely. However, it is easy to see that for $T > 3$, a better strategy is to flip coin 2 initially and then, upon observing a tail, switch to coin 1 for the remaining time horizon. It is not difficult to show that this strategy is optimal and the improvement over the myopic strategy is $(T-3)/8$.

In this simple example, arm 1 is known (thus no information to be gained from playing it) whereas a single play of arm 2 reveals complete information. The tradeoff between exploitation (playing arm 1) and exploration (playing arm 2) is thus easy to appreciate and simple to quantify. In a general bandit problem, however, balancing immediate payoff and information is seldom this simple.

1.3 TWO FORMULATIONS: BAYESIAN AND FREQUENTIST

Similar to other problems with unknowns, there are two schools of approaches to formulating the bandit problems. One is the so-called Bayesian point of view in which the unknown values $\{\theta_i\}$ are random variables with given prior distributions. In this case, the learning algorithm aims to learn the *realizations* of $\{\theta_i\}$ by sharpening the prior distributions using each observation. The performance of the algorithm is measured by averaging over all possible realizations of $\{\theta_i\}$ under the given prior distributions. The other is the frequentist point of view in which $\{\theta_i\}$ are deterministic unknown parameters. Under the frequentist formulation, the problem is a special class of reinforcement learning, where the decision maker learns by coupling actions with observations. Not averaged over all possible values of $\{\theta_i\}$, the performance of a learning algorithm is inherently dependent on the specific values of the unknown parameters $\{\theta_i\}$.

To contrast these two approaches, consider again the archetypal problem of flipping two coins with unknown biases. The Bayesian approach can be interpreted as addressing the following scenario: we are given two buckets of coins, each containing a known composition of coins with different biases; we randomly selection one coin from each bucket and start flipping the two chosen coins. Under the frequentist approach, however, we face two coins that we have no prior knowledge about whether or which buckets they may come from.

1.3.1 THE BAYESIAN FRAMEWORK

The first Bandit problem posed by Thompson was studied under the Bayesian approach, where Thompson adopted the uniform distribution as the prior and focused on maximizing the expected number of successes in T trials. Known as Thompson sampling, the solution he proposed has been studied and applied within both Bayesian and non-Bayesian frameworks and continues to receive great attention.

Bellman, 1956 [30] formulated the Bayesian bandit problem as a Markov decision process (MDP) and proposed an optimal solution based on dynamic programming. To see the Bayesian bandit problem as an MDP, one only needs to realize that a sufficient statistic for making optimal decisions at any given time is the posterior distributions of $\{\theta_i\}$ computed from the prior distributions and past observations using the Bayes rule. The resulting MDP formulation is readily seen by treating the posterior distributions as the information state which evolves as a Markov process under a given strategy. The Bayesian bandit model defined as a class of MDP problems, however, represents a much broader family of decision problems than the original problem of sampling unknown processes: the state of the MDP is not necessarily restricted to being *informational*, and the state transitions do not necessarily obey the Bayes rule.

While dynamic programming offers a general technique for obtaining the optimal solution, its computational complexity grows exponentially with the number of arms since the state space of the resulting MDP is the Cartesian product of the state spaces of all arms. A natural question is whether bandit problems, as a special class of MDP, possess sufficient structures that admit simple optimal solutions.

The problem fascinated the research community for decades, while the answer eluded them until early 1970s. A legend, as told by Whittle,[1] 1980, put it this way: “The problem is a classic one; it was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage.”

The problem was eventually solved by British mathematicians, after all. The breakthrough was made by Gittins and Jones in 1972, who first read their result at the European Meeting of Statisticians with proceedings published in 1974 [87]. The result, however, became widely known only after the publication of Gittins’ 1979 paper [88].

[1] This Monty Python-styled story was told by Peter Whittle, a Churchill Professor of Mathematics for Operational Research at the University of Cambridge, in the discussion section of Gittins’s 1979 paper [88].

Referred to as the Gittins index theorem, the result states that the optimal policy to the bandit problem exhibits a structure of strong decomposability: a priority index, depending solely on the characterizations of each individual arm, can be attached to each state of each arm, and playing the arm with the currently greatest index is optimal. Such an index policy, by fully decoupling the arms, reduces an N-dimensional problem to N independent one-dimensional problems, consequently reducing the computational complexity from an exponential order to a linear order in the number of arms.

This breakthrough generated a flurry of activities in two directions, as fittingly described by Whittle as exploitation—seeking proofs that provide additional insights and different perspectives—and exploration—examining to what extent this class of MDP can be generalized while preserving the index theorem. We highlight below a couple of celebrated results, leaving the detailed discussion to Chapters 2 and 3.

One notable result in the pursuit of a better understanding was the proof by Weber in 1992 [208], which is "expressible in a single paragraph of verbal reasoning" as described by Whittle in his foreword to the second edition of Gittins' monograph with the addition of two coauthors, Glazebrook and Weber, 2011, [90]. Such a result makes one appreciate better the quote by Richard Feynman: "If a topic cannot be explained in a freshman lecture, it is not yet fully understood."

Along the direction of exploration, a significant extension is the restless bandit problems posed and studied by Whittle in 1988 [212]. This class of more general MDP problems allows for system dynamics that cannot be directly controlled. More specifically, the state of an arm may continue to evolve when it is not engaged. In the original bandit model for sampling unknown processes, a passive arm, with no new observations to update its information state, remains frozen. Under the more general MDP model, however, the state of an arm may represent a certain physical state that continues to evolve regardless of the decision maker's action. For instance, the quality of a communication channel may change even when it is not accessed; queues continue to grow due to new arrivals even when they are not served; targets continue to move even when they are not monitored.

Gittins index policy loses its optimality for restless bandit problems. Based on a Lagrangian relaxation of the problem, Whittle proposed an index policy, known as the Whittle index. It was later shown by Weber and Weiss in 1990 [206] that the Whittle index policy is asymptotically (as the number of arms approaching infinity) optimal under certain conditions. In the finite regime, the optimality of whittle index has been established in a number of special cases motivated by specific engineering applications, and the strong performance of whittle index has been observed in extensive numerical studies. The optimal solution to a general restless bandit problem remains open.

1.3.2 THE FREQUENTIST FRAMEWORK

Within the frequentist framework, since the performance of a strategy in general depends on the unknown parameters $\{\theta_i\}$, an immediate question is how to compare strategies given that, in general, there does not exist a strategy that dominates all other strategies for all values of $\{\theta_i\}$.

Two approaches have been considered. The first one, referred to as the uniform-dominance approach, restricts the set of admissible strategies to *uniformly good* policies: policies that offer "good" (to be made precise later) performance for all $\{\theta_i\}$. This excludes from consideration those heavily biased strategies, such as one that always plays arm 1 and offers the best possible return when θ_1 is indeed the largest. Among thus defined admissible strategies, it is then possible to compare performance as a function of $\{\theta_i\}$ and define the associated optimality.

The second approach is the minimax approach: every strategy is admissible, but the performance of a strategy is measured against the worst possible $\{\theta_i\}$ specific to this strategy and the horizon length T. The worst-case performance of a strategy no longer depends on $\{\theta_i\}$. Meaningful comparison of all strategies can be carried out, nontrivial bounds on the best achievable performance can be characterized, and optimality can be defined. This approach can also be viewed through the lens of a two-player zero-sum game where $\{\theta_i\}$ are chosen by an opponent.

Under both approaches, the main questions of concern are whether the maximum return enjoyed by an omniscient player (the oracle, so to speak) with prior knowledge of $\{\theta_i\}$ can be approached via online learning, and if yes, what the fastest rate of convergence to this maximum return would be. With $\{\theta_i\}$ known, the oracle would obviously be playing the arm with the largest mean value at all time and obtain the maximum expected reward of $\max_i\{\theta_i\}$ per play.

The first attempt at answering the above questions was made by Robbins in 1952 [168], in which he considered a two-armed bandit problem and provided a positive answer to the first question. To address the second question on the convergence rate, in 1985, Lai and Robbins [126] proposed a finer performance measure of *regret*, defined as the expected cumulative loss over the entire horizon of length T with respect to the oracle. All online learning strategies with a regret growing sublinearly in T have a diminishing reward loss per play and asymptotically achieve the maximum average reward of $\max_i\{\theta_i\}$ as T approaches infinity. The specific sublinear regret growth rates, however, differentiate them in their effectiveness of learning.

In this much celebrated work, Lai and Robbins, 1985 [126] took the uniform-dominance approach by restricting admissible strategies to consistent policies. They showed that the minimum regret feasible among consistent policies has a logarithmic order in T and constructed asymptotically optimal policies for several reward distributions. Following this seminal work, simpler sample-mean based index-type policies were developed by Agrawal, 1995 [4] and Auer, Cesa-Bianchi, and Fischer, 2002 [23] that achieve the optimal logarithmic regret order under different conditions on the reward distributions. A couple of open-loop strategies with a predetermined control of the exploration-exploitation tradeoff have also been shown to be sufficient to offer order optimality. These simple strategies, less adaptive to random observations, are recently shown to have advantages over fully adaptive strategies when risk is factored into the performance measure (see Section 5.4.1).

Early results under the minimax approach can be traced back to the work by Vogel in 1960, who studied a two-armed bandit with Bernoulli rewards. It was shown by Vogel, 1960b [204] that the minimum regret growth rate under the minimax approach is of the order $\Omega(\sqrt{T})$. This fundamental result, established more than two decades earlier than its counterpart under the uniform-dominance approach by Lai and Robbins in 1985 seems to have largely escaped the attention of the research community. One notable exception is the detailed and more general coverage in Chapter 9 of the book by Berry and Fristedt, 1985 [33]. In most of the literature, however, the $\Omega(\sqrt{T})$ lower bound on the minimax regret is credited to much later studies appeared in the 2000s.

Stemming from these foundational results, there is an extensive and fast-growing literature, in both theory and applications, on the frequentist bandit model and its variant of adversarial bandits. Chapters 4 and 5 give a detailed coverage of both classical results and recent development.

1.4 NOTATION

Notations used in this book are relatively standard. Random variables and their realizations are denoted by capital and lowercase letters, respectively. Vectors and matrixes are in bold face. Sets are in script style, with $\mathcal{R}$ denoting the set of real numbers.

Probability measure and expectation are denoted by $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$, respectively. The indicator function is denoted by $\mathbb{I}[\cdot]$.

The technical treatment mostly deals with discrete time, in which case, time starts at $t = 1$. For continuous time, the origin is set at $t = 0$.

We follow the Bachmann–Landau notation of $O(\cdot)$ and $\Omega(\cdot)$ for describing the limiting behavior of a function. Throughout the book, the terms strategy, policy, and algorithm are used interchangeably.

CHAPTER 2

Bayesian Bandit Model and Gittins Index

In this chapter, we formulate the Bayesian bandit problem as a Markov decision process (MDP). We then devote our attention to the definition, interpretations, and computation of the Gittins index and the index theorem establishing the optimality of the Gittins index policy.

2.1 MARKOV DECISION PROCESSES

We briefly review basic concepts of MDP. We restrict our attention to discrete-time MDPs with countable state space to illustrate the basic formulation and solution techniques. Readers are referred to texts by Ross, 1995 [171] and Puterman, 2005 [164] for more details.

2.1.1 POLICY AND THE VALUE OF A POLICY

The state $S(t) \in \mathcal{S}$ of a process is observed at discrete time instants $t = 1, 2, \ldots$, where $\mathcal{S}$ is referred to as the state space. Based on the observed state, an action $a \in \mathcal{A}$ is chosen from the action space $\mathcal{A}$. If the process is in state $S(t) = s$ and action a is taken, then the following two events occur.

- A reward $r(s, a)$ is obtained.
- The process transits to state $s' \in \mathcal{S}$ with probability $p(s, s'; a)$.

This decision process is called an MDP. The objective is to design a policy that governs the sequential selection of actions to optimize a certain form of the rewards accrued over a given time horizon.

A *policy* π is given by a sequence of *decision rules* d_t, one for each time t:

$$\pi = (d_1, d_2, \ldots). \tag{2.1}$$

Based on the information used for selecting the actions, a decision rule can be either Markov or history dependent: a Markov decision rule d_t depends on only the state of the process at time t; a history-dependent decision rule uses the entire history of the process (both the states and the actions) up to t. Based on how the chosen action is specified, a decision rule can be either deterministic or randomized. The former returns a specific action to be taken. The latter determines a probability distribution on $\mathcal{A}$ based on which a random action will be drawn. Thus, a Markov deterministic decision rule is a mapping from $\mathcal{S}$ to $\mathcal{A}$. A Markov randomized decision rule is a mapping from $\mathcal{S}$ to the set of probability distributions on $\mathcal{A}$. A policy is Markov if d_t is Markov for all t. A policy is deterministic if d_t is deterministic for all t. A policy is *stationary* if it is Markov and employs the same decision rule for all t. Under a stationary policy, the sequence of the induced states $\{S(t)\}_{t\geq 1}$ forms a homogeneous Markov process. The sequence of the states and the rewards $\{S(t),\ r(S(t), d_t(S(t))\}_{t\geq 1}$ induced by a Markov policy $\pi = (d_1, d_2, \ldots)$ is a *Markov reward process*.

We are interested in policies that are optimal in some sense. Commonly adopted optimality criteria include the total expected reward over a finite horizon of length T, the total expected discounted reward over an infinite horizon, and the average reward over an infinite horizon. Under each of these three criteria, the *value* of a policy π for an initial state s is given as follows.

$$\text{Total reward criterion:} \quad V_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=1}^{T} r(S(t), a(t)) \;\middle|\; S(1) = s\right], \tag{2.2}$$

$$\text{Discounted reward criterion:} \quad V_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=1}^{\infty} \beta^{t-1} r(S(t), a(t)) \;\middle|\; S(1) = s\right], \tag{2.3}$$

$$\text{Average reward criterion:} \quad V_\pi(s) = \liminf_{T\to\infty} \frac{1}{T}\, \mathbb{E}_\pi\left[\sum_{t=1}^{T} r(S(t), a(t)) \;\middle|\; S(1) = s\right], \tag{2.4}$$

where $a(t)$ is the action taken at time t under π, $\mathbb{E}_\pi$ represents the conditional expectation given that policy π is employed, and β $(0 < \beta < 1)$ is the discount factor. The discounted reward criterion is economically motivated: a reward to be earned in the future is less valuable than one earned at the present time.

With the value of a policy defined, the value of an MDP with an initial state s is given by

$$V(s) = \sup_\pi V_\pi(s). \tag{2.5}$$

A policy π^* is optimal if, for all $s \in \mathcal{S}$,

$$V_{\pi^*}(s) = V(s). \tag{2.6}$$

We shall assume the existence of an optimal policy (see, for example, Puterman, 2005 [164] on sufficient conditions that guarantee the existence of an optimal policy). It is known that if an optimal policy exists, then there exists an optimal policy which is Markov and deterministic. The sufficiency of Markov policies for achieving optimality can be seen from the fact that the immediate rewards and the transition probabilities depend on the history only through the current state of the process. The sufficiency of deterministic policies can be easily understood since the return of a random action is a weighted average over the returns of all actions, which cannot exceed the maximum return achieved by a specific action. Based on this result, we can restrict our attention to Markov and deterministic policies.

2.1.2 OPTIMALITY EQUATION AND DYNAMIC PROGRAMMING

It is instructive to consider first the finite-horizon problem in (2.2). We present a technique, known as backward induction, for solving for the optimal policy π^* recursively in time, starting from the last decision period T.

In decision period t, given the current state $S(t) = s$, irrespective of past actions and rewards, the optimal decision rules for the remaining time horizon are those that maximize the expected total reward summed over t to T. In other words, we face a "new" finite-horizon MDP with a horizon length of $T - t + 1$ and an initial state s.

Let $V_n(s)$ $(n = 1, \ldots, T)$ denote the value of the n-stage MDP with an initial state s. We can solve for the value function $V_T(s)$ and the corresponding optimal policy $\pi^* = (d_1^*, \ldots, d_T^*)$ recursively in terms of the horizon length starting from $n = 1$. This can be viewed as going backward in time starting from the last decision period T which faces a simple one-stage decision problem. Specifically, at $t = T$, we readily have, for all $s \in \mathcal{S}$,

$$V_1(s) = \max_{a \in \mathcal{A}} r(s, a). \tag{2.7}$$

The optimal decision rule for this stage maps from each state $s \in \mathcal{S}$ to an action $a^*(s)$ that achieves the above maximum value:

$$a^*(s) = \arg\max_{a \in \mathcal{A}} r(s, a). \tag{2.8}$$

Now consider an n-stage problem with an initial state s. If action a is taken initially, then a reward of $r(s, a)$ is immediately accrued and the state transits to s' with probability $p(s, s'; a)$. We are then facing an $(n - 1)$-stage problem, from which the maximum total remaining reward is given by $V_{n-1}(s')$ under a realized state transition to s'. Thus, if we take action a initially and then follow the optimal decision rules for the subsequent $(n - 1)$-stage problem, the total return is

$$r(s, a) + \sum_{s' \in \mathcal{S}} p(s, s'; a) V_{n-1}(s').$$

The optimal initial action needs to strike a balance between the above two terms: the resulting immediate reward $r(s, a)$ and the impact on future state evolutions (hence future values) characterized by $\{p(s, s'; a)\}$. The value for the n-stage problem is thus given by

$$V_n(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \sum_{s' \in \mathcal{S}} p(s, s'; a) V_{n-1}(s') \right\}. \tag{2.9}$$

The optimal decision rule can be obtained by finding an action attaining the above maximum value for each $s \in \mathcal{S}$.

Equation 2.9 is known as the *optimality equation* or the dynamic programming equation. The above also gives a numerical technique, referred to as *backward induction*, for solving for the T-stage value function $V_T(s)$ and the optimal policy $\pi^* = (d_1^*, \ldots, d_T^*)$ recursively.

Consider next the criterion of discounted reward given in (2.3). The value function $V(s)$ satisfies, for all $s \in \mathcal{S}$, the following optimality equation:

$$V(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \beta \sum_{s' \in \mathcal{S}} p(s,s';a) V(s') \right\}. \tag{2.10}$$

The above functional equation of $V(s)$ can be understood in a similar way as (2.9). After taking action a under the initial state s at $t = 1$, we face, at $t = 2$, the same infinite-horizon discounted MDP problem, except with a potentially different initial state s' and an additional multiplicative factor β for all rewards (due to a starting point of $t = 2$). The maximum total discounted reward under this initial action a is thus given by the sum of the immediate reward $r(s,a)$ at $t = 1$ and a weighted average of $V(s')$ times β that represents the optimal return from $t = 2$ onward with the weight given by the probability of seeing state s' at $t = 2$. Optimizing over the initial action a gives us $V(s)$.

The following theorem summarizes several important results on MDP with discounted rewards. We assume here the rewards $\{r(s,a)\}$ are bounded for all s and a.

Theorem 2.1 MDP with Discounted Rewards:

1. *There exists an optimal policy which is stationary.*
2. *The value function is the unique solution to the optimality equation* (2.10).
3. *A policy is optimal if and only if, for all $s \in \mathcal{S}$, it chooses an action that achieves the maximum on the right-hand side of* (2.10).

As defined in Section 2.1.1, a stationary policy employs the same decision rule in all decision periods: $\pi = [d, d, \ldots]$. We can thus equate a stationary policy with its decision rule and directly consider a stationary policy π as a mapping from the state space $\mathcal{S}$ to the action space $\mathcal{A}$. Theorem 2.1-1 states that stationary policies suffice for achieving optimality for MDPs under the infinite-horizon discounted-reward criterion.

Several solution techniques exist for solving for or approximate the value $V(s)$ and an optimal policy. We present here the method of *value iteration*, also referred to as *successive approximation*. This method can be viewed as computing the infinite-horizon value $V(s)$ as the limit of the finite-horizon values $V_n(s)$ as the horizon length n tends to infinity. Following (2.7) and (2.9) and incorporating the discounting, we obtain the following iterative algorithm for computing an ϵ-optimal policy π^*_ϵ whose value is within ϵ of $V(s)$ for all $s \in \mathcal{S}$.

The initial value $V_1(s)$ can be set to arbitrary values, not necessarily the value of the one-stage problem as given in Algorithm 2.1. The global convergence of the value iteration algorithm is guaranteed by the fixed point theorem under the contraction mapping given in (2.10) where the contraction is due to the discounting of $\beta < 1$. For details, see, for example, Puterman, 2005 [164].

Algorithm 2.1 Value Iteration

Input: $\epsilon > 0$.

```
1: Initialization: set TERMINATE = 0, $n = 1$, and
       $$V_1(s) = \max_{a \in \mathcal{A}} r(s,a).$$
2: while TERMINATE = 0 do
3:     $n = n + 1$.
4:     Compute, for all $s \in \mathcal{S}$,
       $$V_n(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \beta \sum_{s' \in \mathcal{S}} p(s,s';a) V_{n-1}(s') \right\}.$$
5:     if $\max_s |V_n(s) - V_{n-1}(s)| < \epsilon$ then
6:         TERMINATE = 1.
7:         Compute, for all $s \in \mathcal{S}$,
           $$\pi^*_\epsilon(s) = \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \beta \sum_{s' \in \mathcal{S}} p(s,s';a) V_n(s') \right\}.$$
8:     end if
9: end while
```

2.2 THE BAYESIAN BANDIT MODEL

Gittins and Jones, 1974 [87] studied the following sequential decision model, which generalizes the Bayesian formulation of multi-armed bandit problems.

Definition 2.2 *The Multi-Armed Bandit Model:*
Consider N arms, each with state space $\mathcal{S}_i$ ($i = 1, \ldots, N$). At time $t = 1, 2, \ldots$, based on the observed states $[S_1(t), \ldots, S_N(t)]$ of all arms, one arm is selected for activation. The active arm, say arm i, offers a reward $r_i(S_i(t))$ dependent of its current state and changes state according to a transition law $p_i(s, s')$ ($s, s' \in \mathcal{S}_i$). The states of all passive arms remain frozen. The objective is an arm activation policy that maximizes the expected total discounted reward with a discount factor β ($0 < \beta < 1$).

The MDP formulation of the bandit problem is readily seen. The state space $\mathcal{S}$ is the Cartesian product of the state spaces of all arms:

$$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \ldots \mathcal{S}_N. \tag{2.11}$$

The action space consists of the indexes of the N arms: $\mathcal{A} = \{1, 2, \ldots, N\}$. The reward and transition probabilities under state $[s_1, \ldots, s_N]$ and action a are given by

$$\begin{aligned} r([s_1, \ldots, s_N], a) &= r_a(s_a), & (2.12)\\ p([s_1, \ldots, s_N],\ [s_1', \ldots, s_N']; a) &= \begin{cases} p_a(s_a, s_a') & \text{if } s_i = s_i' \text{ for all } i \neq a \\ 0 & \text{otherwise} \end{cases}. & (2.13) \end{aligned}$$

In the example below, we show that the first bandit problem posed by Thompson, 1933 [190] falls under the above MDP formulation, except that Thompson adopted the total reward criterion over a finite horizon.

Example 2.3 *The First Bandit Problem and Thompson Sampling:*
Consider two Bernoulli-distributed arms with unknown parameters Θ_1 and Θ_2. Suppose that Θ_1 and Θ_2 are independent with a uniform distribution over [0, 1]. The objective is to maximize the expected total reward over a given horizon by sequentially selecting one arm to play at each time.

The MDP formulation of this bandit problem is as follows. At time t, the state $S_i(t)$ of arm i is the posterior distribution of Θ_i given past observations. In other words, the state $S_i(t)$ takes the form of a non-negative real-valued function with support on [0, 1]. In particular, the initial state of each arm is a constant function of 1 on [0, 1]. Given its current state $S_i(t) = f(\theta)$ (a probability density function), the reward and the state transition of an active arm are given as

$$\begin{aligned} r_i(f) &= \mathbb{E}_f[\Theta_i], & (2.14)\\ S_i(t+1) &= \begin{cases} \frac{\theta f(\theta)}{\mathbb{E}_f[\Theta_i]} & \text{with probability } \mathbb{E}_f[\Theta_i] \\ \frac{(1-\theta) f(\theta)}{1-\mathbb{E}_f[\Theta_i]} & \text{with probability } 1 - \mathbb{E}_f[\Theta_i] \end{cases}, & (2.15) \end{aligned}$$

where $\mathbb{E}_f[\Theta_i] = \int_0^1 \theta f(\theta) d\theta$ is the expected value of Θ_i under distribution f. The state transition follows the Bayes rule: the updated posterior distribution takes two possible forms, depending on whether 1 is observed (with probability $\mathbb{E}_f[\Theta_i]$) or 0 is observed (with probability $1 - \mathbb{E}_f[\Theta_i]$) after playing arm i at time t. In other words, for an active arm i, the state transition law is given by

$$\begin{aligned} p_i\left(f(\theta), \frac{\theta f(\theta)}{\mathbb{E}_f[\Theta_i]}\right) &= \mathbb{E}_f[\Theta_i], & (2.16)\\ p_i\left(f(\theta), \frac{(1-\theta) f(\theta)}{1 - \mathbb{E}_f[\Theta_i]}\right) &= 1 - \mathbb{E}_f[\Theta_i]. & (2.17) \end{aligned}$$

While the MDP formulation of the Bayesian bandit problem is clear, the resulting MDP model is computationally prohibitive due to the complexity of the state space. Thompson proposed a heuristic randomized stationary policy, later known as Thompson Sampling. The policy is as follows. At time t, given the current states $(f_1(\theta_1), f_2(\theta_2))$ of the two arms, the probability $q(f_1(\theta_1), f_2(\theta_2))$ that arm 1 is better than arm 2 is computed:

$$\begin{aligned} q(f_1(\theta_1), f_2(\theta_2)) &\overset{\Delta}{=} \mathbb{P}\left[\Theta_1 > \Theta_2 \mid \Theta_1 \sim f_1(\theta_1), \Theta_2 \sim f_2(\theta_2)\right] && (2.18)\\ &= \int_{\theta_1=0}^{1} \int_{\theta_2=0}^{\theta_1} f_1(\theta_1) f_2(\theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2. && (2.19) \end{aligned}$$

A randomized action is then generated that plays arm 1 with probability $q(f_1(\theta_1), f_2(\theta_2))$ and plays arm 2 with probability $1 - q(f_1(\theta_1), f_2(\theta_2))$. An equivalent and computationally simpler implementation is to draw two random values of θ_1 and θ_2 according to their posterior distributions (i.e., their current states) $f_1(\theta_1)$ and $f_2(\theta_2)$, respectively, and then play arm 1 if $\theta_1 > \theta_2$ and play arm 2 otherwise.

Based on the realization of the random action and the resulting random observation, the state of the chosen arm is updated. The next action is then chosen under the same randomized decision rule.

The above example shows that the original bandit problem posed by Thompson is a special, albeit complex, case of the MDP model specified in Definition 2.2. They are special in the sense that the state of each arm is a probability distribution and the state transitions follow Bayes rule. The general model given in Definition 2.2 is now referred to as the multi-armed bandit problem, while the original Bayesian bandit as posed by Thompson is often referred to as the bandit sampling processes (see Gittins, Glazebrook, and Weber, 2011 [90]).

2.3 GITTINS INDEX

Given that the bandit model is a special class of MDP problems, standard methods for solving MDP directly apply. The computational complexity of such methods, being polynomial in the size of the state space, would grow exponentially with N since the state space of the resulting MDP is the Cartesian product of the state spaces of all arms. Whether this special class of MDP exhibits sufficient structures to allow simple optimal solutions was a classical challenge that intrigued the research community for decades. It was until early 1970s that Gittins and Jones revealed an elegant solution, almost 40 years after Thompson posed the first bandit problem.

2.3.1 GITTINS INDEX AND FORWARD INDUCTION

Gittins and Jones, 1974 [87] and Gittins, 1979 [88] showed that an *index policy* is optimal for the bandit model. Specifically, there exists a real-valued function ν_i that assigns an *index* $\nu_i(s)$ to each state s of each arm i, and the optimal policy is simply to compare the indexes of all the arms at their current states and play the one with the greatest index. The state of the chosen arm then evolves. The index of its new state at the next time instant is compared with the other arms (whose states, hence index values, are frozen) to determine the next action.

The optimality of such an index-type policy reveals that there exists a "figure of merit" that fully captures the value of playing a particular arm at a particular state. What is remarkable is that the index function ν_i of arm i depends solely on the Markov reward process $\{p_i(s,s'),\ r_i(s)\}_{s,s'\in\mathcal{S}_i}$ that defines arm i. That is, in computing the optimal policy, the N arms are decoupled and can be treated separately. The computational complexity is thus reduced from being exponential to being linear with N.

This index, known as Gittins index, has the following form:

$$\nu_i(s) = \max_{\tau\geq 1} \frac{\mathbb{E}\left[\sum_{t=1}^{\tau} \beta^{t-1} r_i(S_i(t)) \mid S_i(1)=s\right]}{\mathbb{E}\left[\sum_{t=1}^{\tau} \beta^{t-1} \mid S_i(1)=s\right]}, \tag{2.20}$$

where τ is a *stopping time*.

To gain an intuitive understanding of the above expression, let us first ignore the maximization over τ and set $\tau = 1$. In this case, the right-hand side of (2.20) becomes $r_i(s)$, the immediate reward obtained by playing arm i at its current state s. If we define the index in this way for all states of all arms, the resulting index policy is simply a myopic policy that considers only the present with no regard of the future.

Myopic policies are suboptimal in general, since a state with a low immediate reward but a high probability to transit to highly rewarding states in the future may have more to offer. A straightforward improvement to the myopic policy is the so-called T-step-look-ahead policies by defining an index with $\tau = T$ in the right-hand side[1] of (2.20). As T increases, the performance of a T-step-look-ahead policy improves, but at the cost of increasing computational complexity (often exponential in T).

A further improvement is to look ahead to an optimally chosen random time τ dependent on the evolution of the state rather than a pre-fixed T steps into the future. More specifically, τ is a stopping time defined on the Markov process of the arm states $\{S_i(t)\}_{t\geq 1}$. An example stopping time $\tau_{\mathcal{Q}}$ is the *hitting time* of a subset $\mathcal{Q}$ of states ($\mathcal{Q} \subset \mathcal{S}_i$), that is, the first time the Markov chain enters the set $\mathcal{Q}$ (referred to as the stopping set):

$$\tau_{\mathcal{Q}} \overset{\Delta}{=} \min\{t : S_i(t) \in \mathcal{Q}\}. \tag{2.21}$$

As will be made clear later, it is sufficient to consider only this type of stopping times (for all possible stopping sets $\mathcal{Q} \subset \mathcal{S}_i$) in the maximization of (2.20). This extension to T-step-look-ahead is referred to as *forward induction*, in contrast to the backward induction discussed in

[1] The commonly adopted definition of a T-step-look-ahead policy corresponds to an index measuring the expected total discounted reward accrued over T decision periods, i.e., the numerator of the right-hand side of (2.20). For a deterministic T, however, the denominator in (2.20) is a constant independent of i and s, thus inconsequential in action selection based on the index.

Section 2.1.2. The numerator of the right-hand side of (2.20) is the total discounted reward obtained up to the stopping time τ, and the denominator is the total discounted time to reach τ from an initial state s. The index $\nu_i(s)$ can thus be interpreted as the *maximum rate* (by choosing the optimal stopping time) at which rewards can be accrued from arm i starting at state s. This notion of maximum equivalent reward rate will be made more rigorous in the next subsection.

The following example, based on one in the monograph by Gittins, Glazebrook, and Weber, 2011 [90], is perhaps the most illuminating in understanding the index form given in (2.20).

Example 2.4 Consider N biased coins, each with a known bias θ_i $(i = 1, \dots, N)$. We are allowed to toss each coin once, and receive a unit reward for each head tossed. In what order should the coins be tossed in order to maximize the expected total discounted reward with a discount factor β $(0 < \beta < 1)$?

The answer is rather obvious. Without discounting, every ordering of coin tossing gives the same expected total reward. With discounting, the coins should be tossed in decreasing order of θ_i to minimize reward loss due to discounting.

Now consider N stacks of coins, each consisting of an infinite number of biased coins. Let $\theta_{i,j}$ denote the known bias of the jth coin in the ith stack. At each time, we choose a stack, toss the coin at the top, and remove it from the stack. The objective remains the same: to maximize the expected total discounted reward.

We can quickly recognize this problem as a bandit problem with each stack of coins being an arm and the bias of the coin currently at the top of the stack being the current state of the corresponding arm. This is, however, a much simpler version of the problem in the sense that the state of an active arm evolves deterministically. More specifically, the reward process associated with each arm/stack i is a known deterministic sequence $\{\theta_{i,j}\}_{j \geq 1}$.

Consider first that for each stack i, the coin biases are monotonically decreasing ($\theta_{i,j} \geq \theta_{i,j+1}$ for all j). The optimal strategy is a myopic one: flip the top coin with the greatest bias among the N stacks. Such a strategy interleaves the given N monotonic reward sequences into a single monotonic sequence, thus minimizing reward loss due to discounting. This can be readily seen by drawing a contradiction: if the resulting sequence of rewards is not monotonically decreasing, then at least one greater value of reward is more heavily discounted in place of a smaller reward, resulting in a smaller total discounted reward.

If, however, the coin biases are not monotone, the myopic strategy is in general suboptimal, and the optimal policy is not immediately obvious. Applying the Gittins index in (2.20), we obtain the following strategy. For each stack i, we compute the following index (for simplicity, we relabel coins from top to bottom starting at 1 after each coin removal):

$$\nu_i = \max_{\tau \geq 1} \frac{\sum_{j=1}^{\tau} \beta^{j-1} \theta_{i,j}}{\sum_{j=1}^{\tau} \beta^{j-1}}, \tag{2.22}$$

where the maximization is over all positive integers. Let τ_i^* denote the positive integer that achieves the maximum of the right-hand side above. The index ν_i can be seen as the maximum reward rate achieved by tossing, consecutively, the top τ_i^* coins in the ith stack. The Gittins index policy is to choose the stack, say k, whose index ν_k is the greatest, toss and remove the top τ_k^* coins in this stack, recompute the index for stack k, and repeat the process.

One may immediately notice a discrepancy in the implementation of Gittins index policy described here and that given at the beginning of this subsection. Instead of recomputing and recomparing the indexes after each coin toss, the top τ_k^* coins in the currently chosen stack are tossed consecutively without worrying that the index of this stack may drop below those of other stacks after some of these τ_k^* coins have been tossed and removed. We can show that these two implementations are equivalent. This is due to a monotone property of the Gittins index. In particular, all the $\tau_k^* - 1$ coins immediately following the top coin in stack k have indexes no smaller than that of the top coin. This can be intuitively explained as follows. Had a coin among these $\tau_k^* - 1$ coins had a smaller index (i.e., a lower reward rate), a higher reward rate of the top coin would have been achieved by stopping before tossing this "inferior" coin. This contradicts with the definition of the index ν_k and the optimal stopping time τ_k^*. This monotone property associated with the Gittins index, which also plays a central role in developing numerical algorithms for computing Gittins index, will be made precise in Property 2.5 in Section 2.3.3.

2.3.2 INTERPRETATIONS OF GITTINS INDEX

Before formally stating and proving the optimality of the Gittins index policy, we discuss below several interpretations of the Gittins index to gain insights from different perspectives. Each of these interpretations led to a major proof of the index theorem in the literature, documenting a collective pursuit of understanding spanning multiple decades since Gittins's original proposal of the index theorem.

Calibration with a standard arm: The role of the index in an index-type policy is to provide a comparative measure for deciding which arm to activate. One way to compare the arms without considering them jointly is to compare each arm with a standard one that offers rewards at a constant rate. This standard arm serves the role of calibration.

Consider arm i and a standard arm that offers a constant reward λ each time it is activated. At each time, we may choose to play either arm i or the standard arm. The objective is to maximize the expected total discounted reward. This decision problem is often referred to as the 1.5-arm bandit problem.

The above problem is an MDP with state space $\mathcal{S}_i$. Based on Theorem 2.1, there exists an optimal policy which is stationary (in addition to being deterministic and Markov). Such an optimal policy, being a mapping from $\mathcal{S}_i$ to the two possible actions, partitions the state space $\mathcal{S}_i$ into two disjoint subsets. Let $\mathcal{Q}_i^*(\lambda)$ denote the set of states in which playing the standard arm is optimal. The set $\mathcal{S}_i \setminus \mathcal{Q}_i^*(\lambda)$ thus contains states in which playing arm i is optimal.

We first note that the optimal policy for choosing which arm to play can also be viewed as an optimal stopping rule for quitting playing arm i. The reason is that once the state of arm i enters $\mathcal{Q}_i^*(\lambda)$ for which the standard arm is chosen, the state of arm i stays frozen thus remains in $\mathcal{Q}_i^*(\lambda)$. Consequently, the standard arm will be chosen for the rest of the time horizon. Thus, the optimal policy specifies an optimal stopping time for playing arm i, which is given by the hitting time of $\mathcal{Q}_i^*(\lambda)$. The dependency of the optimal stopping set $\mathcal{Q}_i^*(\lambda)$ on λ is quite obvious: when the reward rate λ of the standard arm is sufficiently small ($\lambda \to -\infty$), it is never optimal to play the standard arm, and $\mathcal{Q}_i^*(\lambda) = \emptyset$; when λ is sufficiently large, we should always play the standard arm, and $\mathcal{Q}_i^*(\lambda) = \mathcal{S}_i$.

Consider a specific initial state $S_i(1) = s$ of arm i. A critical value of λ is given by the case when it is equally optimal to play the standard arm all through (i.e., to include s in $\mathcal{Q}_i^*(\lambda)$) or to play arm i initially until an optimal stopping time given by the hitting time $\tau_{\mathcal{Q}_i^*(\lambda)}$ with s excluded from $\mathcal{Q}_i^*(\lambda)$. In other words, playing arm i starting from the initial state s up to $\tau_{\mathcal{Q}_i^*(\lambda)}$ produces the same expected total discounted reward as a standard arm with a constant reward rate λ. This critical value of λ is thus an obvious candidate for the index of s.

Next we show that this definition of the index leads to the characterization given in (2.20). At the critical value of λ, we have, by definition,

$$\frac{\lambda}{1-\beta} = \max_{\tau \geq 1} \mathbb{E}\left[\sum_{t=1}^{\tau} \beta^{t-1} r_i(S_i(t)) + \beta^{\tau} \frac{\lambda}{1-\beta} \;\middle|\; S_i(1) = s\right], \tag{2.23}$$

where the left-hand side is the total discounted reward for playing the standard arm all through, and the right-hand side is the total discounted reward for playing arm i initially until an optimal stopping time and then switching to the standard arm. Equivalently, for any stopping time $\tau > 1$, we have

$$\frac{\lambda}{1-\beta} \geq \mathbb{E}\left[\sum_{t=1}^{\tau} \beta^{t-1} r_i(S_i(t)) + \beta^{\tau} \frac{\lambda}{1-\beta} \;\middle|\; S_i(1) = s\right], \tag{2.24}$$

which leads to

$$\lambda \geq \frac{\mathbb{E}\left[\sum_{t=1}^{\tau} \beta^{t-1} r_i(S_i(t)) \mid S_i(1) = s\right]}{\mathbb{E}\left[1-\beta^{\tau} \mid S_i(1) = s\right]/(1-\beta)} \tag{2.25}$$

with equality for the optimal stopping time. The characterization of the Gittins index given in (2.20) thus follows.

Calibrating arm i by a standard arm shows that the index $\nu_i(s)$ given in (2.20) represents the *maximum equivalent constant reward rate* offered by arm i with an initial state s. It also shows that it suffices to consider in (2.20) only the hitting times of all possible subsets of $\mathcal{S}_i$ in the maximization over all stopping times τ due to the sufficiency of stationary policies for the 1.5-arm bandit problem. Recall that $\mathcal{Q}_i^*(\lambda)$ denotes the set of states of arm i in which playing the standard arm with reward rate λ is optimal. The stopping time τ that attains the maximum in (2.20) is thus the hitting time of set $\mathcal{Q}_i^*(\nu_i(s))$, where we adopt the convention that arm i is chosen over the standard arm when the two actions are equally optimal (thus s is excluded from $\mathcal{Q}_i^*(\nu_i(s))$ and the constraint of $\tau > 1$ in the right-hand side of (2.20) is satisfied). We simplify the notation of $\mathcal{Q}_i^*(\nu_i(s))$ to $\mathcal{Q}_i^*(s)$ for the stopping set that attains the Gittins index of state s of arm i.

The 1.5-arm bandit problem and the idea of calibration using a known arm trace back to the work by Bradt, Johnson, and Karlin, 1956 [43], who considered the problem of playing an unknown Bernoulli arm against a known one. The finite-horizon total-reward criterion was adopted. They showed that the optimal action at each time is determined by comparing the success rate of the known arm with a critical value attached to the posterior probability of success (i.e., the state) of the unknown arm. This critical value, however, is also a function of the remaining time horizon due to the finite-horizon formulation. Bellman, 1956 [30], considered the same problem under the infinite-horizon discounted-reward criterion, adopting a beta-distributed prior for the unknown arm. The existence of an index form that characterizes the optimal policy was established using the technique of value iteration (although the term "index" was not introduced there). The analytical form of the index, however, was not obtained, and Bellman remarked in the conclusion that "it seems to be a very difficult problem..."

Gittins index as the fair charge: In the proof of the index theorem given by Weber, 1992 [208] the Gittins index was interpreted as a fair charge for playing an arm.

Consider a single arm i. At each decision time, if we decide to play the arm (and collect the reward offered by the arm), a fixed charge λ, referred to as the *prevailing charge* must be paid. If the charge λ is too great, it will not be worthwhile to play the arm at all; the optimal profit we can gain is 0. If the charge is sufficiently small, the optimal strategy would be to play arm i until an optimal stopping time for a positive profit.

Define the *fair charge* as the critical value of the prevailing charge for which it is equally optimal to not play at all and play until an optimal stopping time. In other words, under the fair charge, the best possible outcome is to break even, and the game thus defined is a fair game. Consider a given initial state s of the arm. Let $\zeta_i(s)$ denote the fair charge for state s. The total discounted reward obtained from playing the arm up to a stopping time τ is upper bounded by the total discounted charges paid:

$$\zeta_i(s)\mathbb{E}\left[\sum_{t=1}^{\tau}\beta^{t-1} \mid S_i(1) = s\right] \geq \mathbb{E}\left[\sum_{t=1}^{\tau}\beta^{t-1}r_i(S_i(t)) \mid S_i(1) = s\right], \tag{2.26}$$

where equality holds for the optimal stopping time (i.e., when we play optimally and break even). The same characterization of the index in (2.20) thus follows.

It is not difficult to see the equivalence between the 1.5-arm bandit problem and this single-arm bandit with a prevailing charge. If we subtract a constant λ from the rewards offered by the standard arm and by arm i at every state, the 1.5-arm bandit problem remains mathematically the same. In this case, the reward $r_i(s) - \lambda$ of playing arm i at any state s can be considered as the profit after paying a prevailing charge of λ, and playing the standard arm, now offering zero reward, can be considered as terminating the game. Viewing the index as the fair charge, however, leads to a proof of the index theorem using simple verbal reasoning as detailed in Section 2.4.

Gittins index as the equitable retirement reward: In Whittle's proof of the index theorem, 1980 [210], he considered a single-arm bandit problem with an option of retiring. At each decision time, we can either play arm i or retire with a lump-sum retirement reward of u. For a given initial state s of arm i, the index of this state is defined as the *equitable retirement reward* for which it is equally optimal to retire immediately or to play the arm until an optimal stopping time and then retire.

The equivalence between the 1.5-arm bandit problem and the single-arm bandit with a retirement option is fairly obvious: playing the standard arm is equivalent to retiring except that the retirement reward is paid over an infinite horizon at a constant rate of λ rather than a lump-sum amount of u paid in full at the time of retirement. It is easy to see that a reward rate of $\lambda = u(1-\beta)$ is equivalent to a lump-sum amount of u. As a result, the index defined as the equitable retirement reward differs from the one in (2.20) by a constant factor of $(1-\beta)$. The resulting index policies, however, are identical, since the actions determined by the indexes depend only on the relative order, not the exact values, of the indexes.

Gittins index as the maximum return with restarting: Katehakis and Veinott, 1987 [112] characterized the index based on a restart-in-state formulation of an arm. This characterization leads to an efficient online algorithm for computing the index, as will be discussed in Section 2.5.

Consider arm i with initial state s. Suppose that at each subsequent decision time $t > 1$, we have the option of setting the arm state back to s, thus restarting the arm process. The problem is to decide whether to restart at each time to maximize the expected total discounted reward.

Let $V_i(s)$ denote the value of this MDP. Note that if the action of restarting is taken at time t, then the future value from that point on would be $V_i(s)$ discounted with β^{t-1}. Optimizing over the restarting time τ, we arrive at

$$V_i(s) = \max_{\tau > 1} \mathbb{E}\left[\sum_{t=1}^{\tau-1} \beta^{t-1} r_i(S_i(t)) + \beta^{\tau-1} V_i(s) \mid S_i(1) = s\right], \tag{2.27}$$

which leads to

$$V_i(s) = \max_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{t=1}^{\tau} \beta^{t-1} r_i(S_i(t)) \mid S_i(1) = s\right]}{1 - \mathbb{E}[\beta^{\tau} \mid S_i(1) = s]} = \frac{\nu_i(s)}{1-\beta}. \tag{2.28}$$

Thus, the index of state s can be defined as the maximum return from arm i with the restart option. This definition leads to the same index value as the equitable retirement reward, with both differing from the original definition in (2.20) by a constant factor of $(1-\beta)$. The advantage of this characterization is that $V_i(s)$, as the value function of an MDP, can be computed via standard solution techniques for MDP. This leads to an algorithm for online computation of the Gittins index as discussed in Section 2.5

2.3.3 THE INDEX PROCESS, LOWER ENVELOP, AND MONOTONICITY OF THE STOPPING SETS

We discuss here a monotone property of the stopping sets that attain the Gittins index. This property leads to a numerical method for computing the index, as will be discussed in Section 2.5, as well as an alternative implementation of the index policy as we have seen in Example 2.4. We then introduce the concepts of the index process and its lower envelope developed by Mandelbaum, 1986 [142] in his proof of the index theorem. These concepts provide an intuitive understanding of the index theorem. In the next section, we will see that the prevailing charge process in Weber's proof of the index theorem is precisely the lower envelope of the index process.

Monotonicity of the stopping sets: The following property characterizes the stopping sets that attain the Gittins index.

Property 2.5 Recall that $\mathcal{Q}_i^*(s)$ denotes the stopping set that attains the Gittins index $\nu_i(s)$ of state s of arm i. We have

$$\left\{s' : \nu_i(s') < \nu_i(s)\right\} \subseteq \mathcal{Q}_i^*(s) \subseteq \left\{s' : \nu_i(s') \leq \nu_i(s)\right\}. \tag{2.29}$$

Property 2.5 states that the optimal stopping time $\tau_i^*(s)$ that attains the index $\nu_i(s)$ of state s is to stop once the arm hits a state with an index smaller than $\nu_i(s)$. We give an intuitive explanation of the above property in place of a formal proof. Based on the derivation of the index $\nu_i(s)$ from the 1.5-arm bandit problem, $\mathcal{Q}_i^*(s)$ is the set of states in which playing the standard arm with a constant reward rate of $\nu_i(s)$ is optimal. It is thus easy to see that all states with maximum equivalent constant reward rates (i.e., their Gittins indexes) smaller than $\nu_i(s)$ should be in $\mathcal{Q}_i^*(s)$. Similarly, all states with indexes greater than $\nu_i(s)$ should not be in $\mathcal{Q}_i^*(s)$. States with the same index value as s, including s itself, can be included or excluded from $\mathcal{Q}_i^*(s)$ without affecting the index value.

A direct consequence of Property 2.5 is a monotone property of the optimal stopping sets in terms of the index values they attain: if $\nu_i(s') \leq \nu_i(s)$, then $\mathcal{Q}_i^*(s') \subseteq \mathcal{Q}_i^*(s)$. This monotone property leads to a numerical method for computing the Gittins index when the state space is finite (see Section 2.5). Another consequence of Property 2.5 is an equivalent implementation of the Gittins index policy as detailed below.

The usual implementation of the Gittins index is to compute the index of the current state of each arm at each given time and play the arm with the greatest index. Property 2.5 shows that the following implementation is identical. At $t = 1$, we identify the arm with the greatest index, say it is arm i with an initial state s. We play arm i until the hitting time of $\mathcal{Q}_i^*(s)$. Based on Property 2.5, this is the first time that the index of arm i drops below the index value $\nu_i(s)$ at $t = 1$. We compute the index value of arm i at this time, compare it with the index values of all other arms (which remain unchanged), and repeat the procedure.

The index process and its lower envelope: Property 2.5 implies that if we observe arm i only at the optimal stopping time $\tau_{Q_i^*(s)}$ that attains the index of the previously observed state s, the indexes of the sampled states form a decreasing sequence. We formalize this statement by introducing the concepts of *index process* and the *lower envelope* of an index process.

The most salient feature of the bandit problem is that the player's actions do not affect the state evolution of any arm; they merely pause the evolution of an arm when it is not played. More specifically, each arm is a fixed Markov reward process, and the decision problem is how to interleave[2] these N independent reward processes into a single reward process. The essence of the problem is to interleave these reward processes in such a way that reward segments with higher rates are collected earlier in time to mitigate discounting. This perspective on the bandit problem is most clearly reflected in Example 2.4 where each arm (i.e., a coin stack) is a fixed deterministic reward process and a policy merely decides, repeatedly, which segments of rewards to collect next.

We then equate each arm i with a fixed stochastic reward process $\{r_i(S_i(t))\}_{t\geq 1}$. The *index process* $\{\nu_i(t)\}_{t\geq 1}$ associated with arm i is the induced stochastic process of the Gittins index values, i.e., $\nu_i(t) = \nu_i(S_i(t))$. The *lower envelope* $\{\underline{\nu}_i(t)\}_{t\geq 1}$ of the index process $\{\nu_i(t)\}_{t\geq 1}$ is given by the running minimum of the index values:

$$\underline{\nu}_i(t) = \min_{k\leq t} \nu_i(k). \tag{2.30}$$

Figure 2.1 illustrates a sample path of an index process (the solid line) and its lower envelope (the dashed line).

It is easy to see that each sample path of the lower envelope, as a running minimum, is piecewise constant and monotonically decreasing. Each constant segment of the lower envelope corresponds to the stopping time attaining the index value represented by this segment. We explain this using the sample path illustration in Figure 2.1. The index value of the initial state s determines the first entry in the index process as well as the lower envelope. In the subsequent three time instants, the arm evolves to states with greater values of the index; the lower envelope thus remains at the constant level of $\nu_i(s)$. At $t = 5$, the state s' has a smaller index. This ends the first constant segment of the lower envelope, which now takes the value of $\nu_i(s') < \nu_i(s)$. Based on Property 2.5, this is also the stopping time for attaining the index value of $\nu_i(s)$. The same argument applies to each constant segment of each sample path of the lower envelope.

The definition of the Gittins index in (2.20) applies to general stochastic reward processes beyond Markovian. The property of the lower envelope process and its relation with the optimal stopping times hold for general reward processes as well. As a result, the index theorem holds without the Markovian assumption on the reward processes, which was first shown by Varaiya, Walrand, and Buyukkoc, 1985 [200], and later crystalized in Mandelbaum's analysis (1986 [142]) within the framework of optimal stopping in a multi-parameter process where

[2] Note that the interleaving of the N reward processes resulting from an arm selection policy is stochastic and sample-path dependent in general.

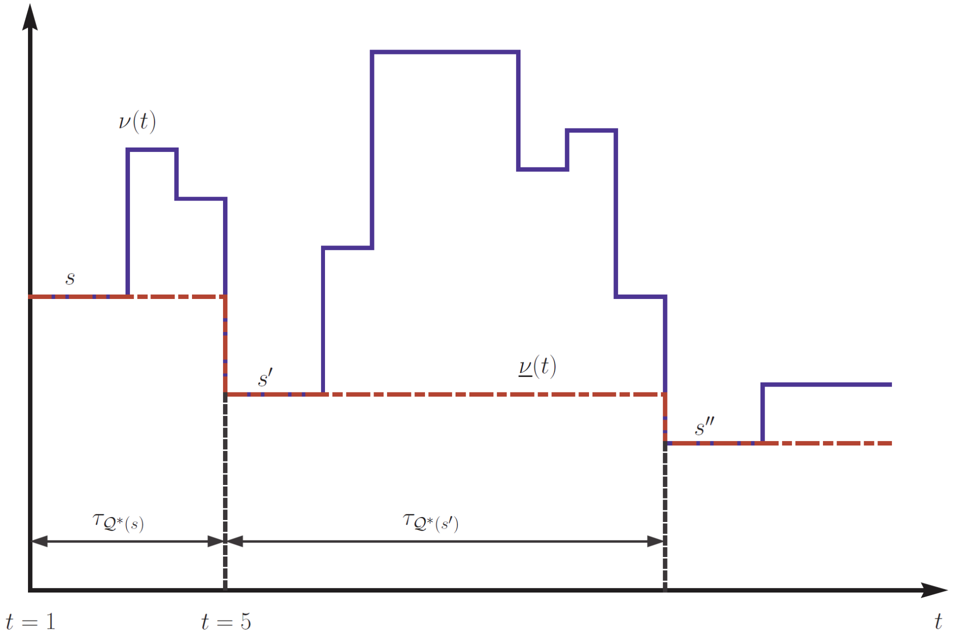


Figure 2.1: The index process and the lower envelope.

time is replaced by a partially ordered set. The proof of the index theorem given in the next section is also from the perspective that casts the bandit problem as interleaving N general stochastic reward processes rather than an MDP.

2.4 OPTIMALITY OF THE GITTINS INDEX POLICY

We now formally state the index theorem and give the proof due to Weber, 1992 [208].

Theorem 2.6 Optimality of Gittins Index Policy:
Consider the multi-armed bandit model defined in Definition 2.2. Let the index $\nu_i(s)$ of every state s of every arm i be defined as in (2.20). The index policy that, at each decision time t for given arm states $\{s_i(t)\}_{i=1}^N$, plays an arm i^* with the greatest index $\nu_{i^*}(s_{i^*}(t)) = \max_{i=1,\ldots,N} \nu_i(s_i(t))$ is optimal.

Since the original work by Gittins and Jones, 1974 [87] and Gittins, 1979 [88], that proved the index theorem based on an interchange argument, a number of proofs using a diverse array of techniques appeared over a span of two decades. The monograph by Gittins, Glazebrook, and Weber, 2011 [90] gave a detailed account of these proofs. We include here the proof by Weber, 1992 [208], based on the fair charge interpretation of the index. We will see that the prevailing charge used in Weber's proof is precisely the lower envelope of the index process as discussed in the previous subsection.

Proof. Consider first a single arm i with an initial state s_0. Recall that the Gittins index $\nu_i(s_0)$ can be interpreted as the fair charge for playing arm i in this initial state (see Section 2.3.2). The prevailing charge for each play of the arm is fixed to the fair charge $\nu_i(s_0)$ of the initial state. The player can choose to play or stop at any given time with the objective of maximizing the expected total discounted profit. The game thus defined is a fair game in which the maximum expected total discounted profit is zero. Assume that the player prefers to continue the game as long as there is no loss. The optimal strategy of the player is to play the arm until it enters a state s' in which the fixed prevailing charge $\nu_i(s_0)$ exceeds the current fair charge $\nu_i(s')$ (i.e., the prevailing charge is too great to continue playing). This optimal stopping time is $\tau_{\mathcal{Q}_i^*(s_0)}$, which yields the maximum profit of zero. Note that if the player stops at a state whose fair charge is greater than the prevailing charge $\nu_i(s_0)$, the player has stopped too early without reaping the profit offered by this advantageous state; a positive loss would incur.

To incentivize continuous play of the game, consider that the prevailing charge is reduced to the fair charge of the current state whenever it is too great for the player to continue. That is, the prevailing charge $\underline{\nu}_i(t)$ at time t is the minimum value of the fair charges of the states seen up to t:

$$\underline{\nu}_i(t) = \min_{k=1,\ldots,t} \nu_i(S_i(k)). \tag{2.31}$$

That is, the prevailing charge process $\{\underline{\nu}_i(t)\}_{t\geq 1}$ is the lower envelope of the index process defined in (2.30), and every sample path of the prevailing charge process is piecewise constant and monotonically decreasing (see Figure 2.1). This modified game remains to be a fair game.

Now suppose there are N arms and at each time the player chooses one of them to play. The prevailing charge of each arm traces the lower envelope of the index process associated with this arm. The resulting game remains to be a fair game, since a strict positive profit from any arm is impossible whereas the maximum expected total discounted profit of zero is achievable (by, for example, playing a single arm forever).

Any policy π for the original bandit problem (i.e., without prevailing charges) is a valid strategy for this fair game, and, when employed, produces at most zero expected profit. The maximum expected total discounted reward in the original bandit problem is thus upper bounded by the maximum expected total discounted prevailing charges that can be collected in this fair game. What remains to be shown is that this upper bound is achieved by the Gittins index policy.

Since the player's actions do not change the state evolution of any arm, we can imagine that the entire sample paths of the states and the associated rewards and prevailing charges are prefixed, although only revealed to the player one value at a time, following the predetermined order, from the arm chosen by the player. Since every sample path of the prevailing charges on each arm is monotonically decreasing, the maximum expected total discounted charges is collected when the player always plays the arm with the greatest prevailing charge, thus interleaving the given N monotonic sequences of prevailing charges into a single monotonic sequence (see Example 2.4 for a deterministic version of the problem and the argument for the optimality of a myopic strategy for interleaving monotone reward sequences). By the construction of the prevailing charges, we arrive at Theorem 2.6. □

2.5 COMPUTING GITTINS INDEX

The Gittins index given in (2.20) may be solved analytically for certain problem instances. In other cases, generic algorithms exist for computing the index numerically.

Various algorithms have been developed for computing the Gittins index. They fall into two categories: offline algorithms and online algorithms. The former compute, in advance, the indices for every possible state of every arm and store them in a lookup table for carrying out the index policy. This approach is mostly suited for bandit processes with a relatively small state space. The latter compute the indices on the fly as each new state is realized. They are particularly efficient in applications that involve a large (potentially infinite) state space, but with a sparse transition matrix promising that only a small fraction of states get realized online. We present in this section one representative algorithm from each category. Chakravorty and Mahajan, 2014 [59] gave an overview of algorithms for computing Gittins index with detailed examples.

It is worth pointing out that the exact values of the Gittins indices are not essential to the implementation of the index policy. What is needed is simply which arm has the greatest index at each given time. There are problems with sufficient structures in which the arm with the greatest index can be identified without explicitly computing the indices. We see such an example in Section 6.1.1 in the context of the Whittle index for a restless bandit problem. There are also iterative algorithms that successively refine the intervals containing the indices until the intervals are disjoint and reveal the ranking of the indices (see, for example, Ben-Israel and Flam, 1990 [32] and Krishnamurthy and Wahlberg, 2009 [121]).

2.5.1 OFFLINE COMPUTATION

The algorithm developed by Varaiya, Walrand, and Buyukkoc, 1985 [200], referred to as the largest-remaining-index algorithm, computes the Gittins indices in decreasing order by exploiting the monotone property of the optimal stopping sets as specified in Property 2.5.

Since arms are decoupled when computing their indices, we consider a single arm and omit the arm index from notations. Suppose that the arm has m states $\mathcal{S} = \{1, 2, \ldots, m\}$ with transition matrix $\mathbf{P}$. Let $(s_1, s_2, \ldots, s_m)$ be a permutation of these m states such that their Gittins indices are in decreasing order:

$$\nu(s_1) \geq \nu(s_2) \geq \ldots \geq \nu(s_m). \tag{2.32}$$

The algorithm proceeds by first identifying the state s_1 and computing its index $\nu(s_1)$. Based on Property 2.5, it is easy to see that the stopping set $\mathcal{Q}^*(s_1)$ can be set to $\mathcal{S}$, the entire state space.

The index of s_1 is thus simply the immediate reward offered by s_1. Since s_1 is the state with the greatest index value (hence the greatest immediate reward), we have

$$\begin{aligned} s_1 &= \arg\max_{s\in\mathcal{S}} r(s), && (2.33)\\ \nu(s_1) &= r(s_1), && (2.34) \end{aligned}$$

where ties are broken arbitrarily.

Suppose that the top $k-1$ largest Gittins indices have been computed and the corresponding states $\{s_1,\ldots,s_{k-1}\}$ identified. The next step is to identify s_k and compute $\nu(s_k)$. It is known from Property 2.5 that $\mathcal{Q}^*(s_k) = \mathcal{S}\setminus\{s_1,\ldots,s_{k-1}\}$. If we use $\tau_{\mathcal{Q}^*(s_k)}$, the hitting time of $\mathcal{Q}^*(s_k)$, as the stopping time in the right-hand side of (2.20) for calculating a hypothetical index for each of the remaining states in $\mathcal{S}\setminus\{s_1,\ldots,s_{k-1}\}$, we know that the thus obtained index for s_k would equal to its Gittins index $\nu(s_k)$ since the optimal stopping time $\tau_{\mathcal{Q}^*(s_k)}$ is used in the calculation. On the other hand, the calculated indices for all other states are no greater than their Gittins indices (due to the potentially suboptimal stopping time used in the calculation), which in turn are no greater than $\nu(s_k)$. We thus obtain $\nu(s_k)$ as the largest value among these hypothetical indices calculated using $\tau_{\mathcal{Q}^*(s_k)}$. Specifically, let $\mathbf{v}_k$ and $\mathbf{w}_k$ denote, respectively, the $m\times 1$ vector of the expected discounted reward and the expected discounted time until $\tau_{\mathcal{Q}^*(s_k)}$ for each initial state $s\in\mathcal{S}$. Based on basic results on calculating expected hitting time of a Markov chain, we have

$$\begin{aligned} \mathbf{v}_k &= \left[\mathbf{I}-\beta\tilde{\mathbf{P}}_k\right]^{-1}\mathbf{r}, && (2.35)\\ \mathbf{w}_k &= \left[\mathbf{I}-\beta\tilde{\mathbf{P}}_k\right]^{-1}\mathbf{1}, && (2.36) \end{aligned}$$

where $\mathbf{r}=\{r(s)\}_{s\in\mathcal{S}}$ is the $m\times 1$ vector of the immediate reward of each state, $\mathbf{1}$ is a vector of all 1's, and $\tilde{\mathbf{P}}_k$ is a modified transition matrix after replacing all columns corresponding to states in $\mathcal{Q}^*(s_k)$ with 0, i.e., for all $i=1,\ldots,m$,

$$\tilde{\mathbf{P}}_k(i,j) = \begin{cases} p(i,j) & \text{if } j\in\{s_1,\ldots,s_{k-1}\} \\ 0 & \text{otherwise} \end{cases}. \qquad (2.37)$$

We then have

$$\begin{aligned} \nu(s_k) &= \max_{j\in\mathcal{S}\setminus\{s_1,\ldots,s_{k-1}\}} \frac{v_k(j)}{w_k(j)}, && (2.38)\\ s_k &= \arg\max_{j\in\mathcal{S}\setminus\{s_1,\ldots,s_{k-1}\}} \frac{v_k(j)}{w_k(j)}, && (2.39) \end{aligned}$$

where $v_k(j)$ and $w_k(j)$ are the elements of $\mathbf{v}_k$ and $\mathbf{w}_k$ that correspond to state j. The procedure continues until all m indices have been computed.

Example 2.7 Consider an arm with three states $\mathcal{S} = \{a, b, c\}$ and the following reward vector and transition matrix:

$$\mathbf{r} = \begin{pmatrix} 2 \\ 5 \\ 3 \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix}. \tag{2.40}$$

Suppose that $\beta = \frac{4}{5}$. Compute the Gittins indices for the three states.

State b has the highest immediate reward of 5. We thus have $\nu(b) = r(b) = 5$, which is the largest index value among the three states.

To identify the state s_2 with the second largest index and its index value, we use the stopping set $\mathcal{Q}^*(s_2) = \mathcal{S} \setminus \{b\} = \{a, c\}$. Rewrite (2.35) in a more intuitive form of

$$\mathbf{v}_k = \mathbf{r} + \beta \tilde{\mathbf{P}}_k \mathbf{v}_k, \tag{2.41}$$

which corresponds to the following three equations easily seen from the modified Markovian state transitions which dictate termination at states a and c:

$$v_2(a) = r(a) + \beta p(a, b) v_2(b), \tag{2.42}$$
$$v_2(b) = r(b) + \beta p(b, b) v_2(b), \tag{2.43}$$
$$v_2(c) = r(c) + \beta p(c, b) v_2(b). \tag{2.44}$$

A similar set of equations for $w_2(a), w_2(b), w_2(c)$ can be written (by replacing the immediate reward in each state with 1 for counting the discounted time rather than rewards until stopping). Solving these equations gives us

$$\mathbf{v}_2 = \begin{pmatrix} 4 \\ 5 \\ 3 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} \frac{7}{5} \\ 1 \\ 1 \end{pmatrix}.$$

State c thus has the second largest index with $\nu(c) = 3$. For the last state a which has the smallest index value, the stopping set is $\{a\}$, which leads to $\nu(a) = 73/44$.

Other offline algorithms for computing Gittins index include variations of the largest-remaining-index by Denardo, Park, and Rothblum, 2007 [73], Sonin, 2008 [182], and Denardo, Feinberg, and Rothblum, 2013 [74], and algorithms based on linear programming by Chen and Katehakis, 1986 [61], and parametric linear programming by Kallenberg, 1986 [106] and Nino-Mora, 2007a [157]. All these algorithms enjoy computational complexity in the order of $O(m^3)$ with differences in the leading constants, where m is the size of the state space.

2.5.2 ONLINE COMPUTATION

The above largest-remaining-index algorithm computes the indexes of *all* states in a descending order of their index values. It does not give an efficient solution when only the indexes of a subset of states are needed.

The interpretation of Gittins index of a state s_0 as the value of an MDP with the option of restarting at s_0 leads to an online computation of indexes for those states seen by the system in real time. Standard MDP solution techniques such as value iteration, policy iteration, and linear programming can be used to compute the value of this two-action MDP. Specifically, the optimality equations of this MDP are as follows. For all $s \in \mathcal{S}$,

$$V(s) = \max \left\{ r(s_0) + \beta \sum_{s' \in \mathcal{S}} p(s_0, s')V(s'),\ r(s) + \beta \sum_{s' \in \mathcal{S}} p(s, s')V(s') \right\}, \tag{2.45}$$

where the first term inside the maximization corresponds to the action of restarting to state s_0 and the second term the action of continuing with the current state s. An ϵ-approximation of the Gittins index of s_0, given by $V(s_0)$ (with the constant factor of $\frac{1}{1-\beta}$ ignored), can be computed using the value iteration algorithm as given in Algorithm 2.1.

2.6 SEMI-MARKOV BANDIT PROCESSES

We have thus far focused on bandit problems where the decision times are discrete and predetermined, which, without loss of generality, can be set to positive integers $t = 1, 2, \ldots$. In this section, we consider bandit problems where the inter-decision times are random and state dependent. We show that the index theorem carries through to such semi-Markov bandit problems.

Definition 2.8 *The Semi-Markov Bandit Model:*
Consider N arms, each with state space $\mathcal{S}_i$ $(i = 1, \ldots, N)$. At each time $t \in \mathcal{R}^+$, only one arm may be active. Time $t_0 = 0$ is the first decision time for choosing one arm to activate. Let t_k denote the kth decision time. Suppose that at time t_k, arm i is in state s and is chosen for activation. Then the following occur.

- A reward $r_i(s)$ is immediately obtained at time t_k.
- The state of arm i at the next decision time t_{k+1} is s' with probability $p_i(s, s')$.
- The inter-decision time $T_i = t_{k+1} - t_k$ of arm i is drawn, independent of previous inter-decision times, according to a distribution[3] $F_i(\cdot \mid s, s')$ that depends on the states s and s' at times t_k and t_{k+1}, i.e.,

$$\Pr\left[T_i \leq x \mid s, s'\right] = F_i(x \mid s, s'). \tag{2.46}$$

[3] We assume the usual technical condition on F_i for semi-Markov decision processes to ensure that an infinite number of state transitions will not occur in finite time (see, for example, Chapter 5 of Ross, 1970 [170] and Section 11.1 of Puterman, 2005 [164]).

The states of all passive arms remain frozen. At the next decision time t_{k+1} determined by T_i drawn from $F_i(\cdot \mid s, s')$, with arm i in state s', a decision on which arm to activate next is made, and the process continues. The objective is to maximize the expected total discounted reward with a discount factor β $(0 < \beta < 1)$.

We give below two examples of semi-Markov bandit model.

Example 2.9 *Continuous-Time Bandit Processes:*
A bandit problem in which each arm, when activated, evolves according to a continuous-time Markov process, is an example of the semi-Markov bandit model, where the decision times are given by the state transition times of activated arms. Let $\mathbf{Q}_i = \{q_i(s, s')\}_{s,s' \in \mathcal{S}_i}$ denote the transition rate matrix of the corresponding continuous-time Markov process associated with arm i. Then the state transition probabilities and the inter-decision times in the resulting semi-Markov bandit model are given by

$$p_i(s, s') = -\frac{q_i(s, s')}{q_i(s, s)}, \quad T_i(s, s') \sim \exp(-q_i(s, s)). \tag{2.47}$$

Note that in this case, the inter-decision time $T_i(s, s')$ depends only on the state s in the current decision period.

Example 2.10 *Bandit Processes with Restricted Switching States:*
Consider a standard discrete-time bandit model as defined in Definition 2.2 with the exception that we cannot switch out of an active arm i until it enters a state in set $\mathcal{W}_i \subseteq \mathcal{S}_i$ $(i = 1, 2, \ldots, N)$. If $\mathcal{W}_i = \mathcal{S}_i$ for all i, we go back to the standard model. If each arm represents a job and $\mathcal{W}_i$ contains the completion state of job i, then this model represents a non-preemptive job scheduling problem (we discuss more the job scheduling problem in Section 3.2.2).

For this problem, the decision time following an activation of an arm (say, i) is when the arm enters a state in $\mathcal{W}_i$. The state transition probabilities in the resulting semi-Markov bandit model are $p_i(s, s') = 0$ for $s' \notin \mathcal{W}_i$. For $s' \in \mathcal{W}_i$, the transition probability $p_i(s, s')$ and the inter-decision time $T_i(s, s')$ are given by the corresponding absorbing probability and absorbing time in the modified Markov process with states in $\mathcal{W}_i$ as absorbing states.

Analogous to (2.20), the index of a semi-Markov bandit process in state s is given by

$$\nu_i(s) = \max_{\tau > 0} \frac{\mathbb{E}\left[\sum_{t_k < \tau} \beta^{t_k} r_i(S_i(t_k)) \,\middle|\, S_i(0) = s\right]}{\mathbb{E}\left[\int_{t=0}^{\tau} \beta^t \, \mathrm{d}t \,\middle|\, S_i(0) = s\right]}, \tag{2.48}$$

where the stopping time τ is constrained to the set of decision times $\{t_1, t_2, \ldots\}$ which are the transition times of the semi-Markov process defined by $\{p_i(s, s')\}_{s,s' \in \mathcal{S}_i}$ and $F_i(\cdot \mid s, s')$.

With relatively minor modifications to the proof given in Chapter 2.4, it can be shown that playing the arm with the greatest index given in (2.48) at each decision time is optimal for semi-Markov bandit problems. A detailed proof can be found in Section 2.8 of Gittins, Glazebrook, and Weber, 2011 [90].

CHAPTER 3

Variants of the Bayesian Bandit Model

It is natural to ask the question that to what extend the bandit model can be generalized without breaking the optimality of the Gittins index policy. Extensive studies have been carried out since Gittins's seminal work, aiming to push the applicability boundary of the index theorem. This chapter is devoted to these efforts. We start by identifying delimiting features of the bandit model that are necessary for the index theorem. We then cover a number of variants of the bandit model that are significant in their technical extensions and general in their applicability. Based on which key feature of the bandit model is being extended, we group these variants into four classes: variations in the action space, in the system dynamics, in the reward structure, and in the performance measure.

3.1 NECESSARY ASSUMPTIONS FOR THE INDEX THEOREM

We list below the key features that delimit the bandit model, followed by a discussion on which modeling assumptions are necessary and which are not for the index theorem.

Modeling Assumptions of the Bandit Model:

▷ On the action space:

A1.1 Only *one* arm can be activated each time.

A1.2 The arms are otherwise *uncontrolled*, that is, the decision maker chooses which arm to operate, but not *how* to operate it.

A1.3 The set of arms are fixed *a priori*, and all arms are available for activation at all times.

▷ On the system dynamics:

A2.1 An active arm changes state according to a homogeneous Markov process.

A2.2 State evolutions are independent across arms.

A2.3 The states of passive arms are frozen.

▷ On the reward structure:

A3.1 An active arm offers a reward dependent only on its own state.

A3.2 Passive arms offer no reward.

A3.3 Arm switching incurs no cost.

▹ On the performance measure:

A4 Expected total (geometrically) discounted reward over an infinite horizon as given in (2.3).

3.1.1 MODELING ASSUMPTIONS ON THE ACTION SPACE

A1.1 is necessary for the index theorem. If $K > 1$ arms can be activated simultaneously, it is in general suboptimal to active the K arms with the top K greatest Gittins indices. This can be easily seen from the following counterexample.

CounterExample 3.1 Consider Example 2.4 with $N = 3$ stacks of coins. The expected reward sequence, starting from the top coin, is $(0.8, 0, 0, \ldots)$ for the first and the second stacks, and $(0.6, 0.96, 0, 0, \ldots)$ for the third stack. At each time, we are allowed to flip $k = 2$ coins simultaneously. The discount factor β is set to 0.5.

It is easy to see that at $t = 1$, the Gittins indices of these three arms are 0.8, 0.8, and 0.72, respectively. If we choose the two arms with the top two greatest Gittins index to activate at each time, we would choose the top coins in the first and the second stacks to flip at $t = 0$, accruing a total expected reward of $0.8 + 0.8 = 1.6$, and then flip the top coin with bias 0.6 in the third stack along with a zero-reward coin from either of the first two stacks at $t = 1$, and finally flip the second coin with bias 0.96 in the third stack at $t = 3$. There is no reward to be earned afterward. The total discounted reward is $1.6 + 0.6\beta + 0.96\beta^2 = 2.14$.

If, on the other hand, we flip the top coins in the first and the third stacks at $t = 1$ and then the top coin in the second stack along with the second coin in the third stack at $t = 2$, we would have obtained a total discounted reward of $(0.8 + 0.6) + (0.8 + 0.96)\beta = 2.28$. It is easy to see that this is the optimal strategy.

Rather than activating the arms with the top k greatest Gittins indices, one might wonder whether there are other ways of extending the index policy that preserve the optimality in the case of simultaneous multi-arm activation. For instance, we may consider to define each subset of k arms as a super-arm with a k-dimensional state space and obtain its index in a similar way as (2.20). The hope here is that such an index would capture the maximum rate of the rewards jointly offered by k arms and A1.1 is satisfied in terms of super-arms. Unfortunately, for such a bandit model, A2.2 and A2.3 no longer hold. In particular, a passive super-arm may change state if it shares common arms with the activated super-arm. As discussed below, both A2.2 and A2.3 are necessary for the optimality of the index policy. The general structure of the optimal policy in the case of simultaneous multi-arm activation remains open.

A1.2 is also necessary. When there exist multiple ways to activate an arm, each resulting in different rewards and associated with different state transition probabilities, the problem is referred to as the *bandit superprocess model*. This variant in general does not admit an index structure. Sufficient conditions, albeit restrictive, have been established for the optimality of the index policy. We discuss this variant in detail in Section 3.2.1.

There are two general directions to relax A1.3. One is to allow precedence constraints on available arms: an arm may not be available for activation until several other arms reach certain states. Such precedence constraints are common in job scheduling where a job cannot start until the completion of several other jobs. The other direction is to allow arrivals of new arms throughout the decision horizon. We discuss in Section 3.2.2 classes of precedence constraints and arrival processes of new arms that preserve the optimality of the index policy.

3.1.2 MODELING ASSUMPTIONS ON THE SYSTEM DYNAMICS

A2.1 on the Markovian dynamics of state evolution under activation is crucial to the MDP formulation of the problem. It is, however, not necessary for the optimality of the index policy as discussed in Section 2.3.3. While the Markovian assumption of the arm evolutions is not necessary for the index theorem, it is crucial to the numerical algorithms for computing the index as seen in both the offline and online algorithms discussed in Section 2.5. For arbitrary reward processes, the index, given by the supremum value over all stopping times on an arbitrary stochastic process (see (2.20)), is generally intractable to compute even numerically.

A2.2 on the statistical independence across arms is necessary for the optimality of the index policy, as intuition suggests. Problems involving localized arm dependencies can be translated into the bandit superprocess model. Consider, for example, arm 1 and 2 are dependent in an N-arm bandit problem. We can treat arm 1 and 2 jointly as a superprocess with two control actions (playing arm 1 or 2 when this superprocess is activated). The value of such a formulation, however, depends on how local the dependencies are across arms and whether the resulting superprocesses admit an index rule as an optimal solution.

A2.3 is necessary for the optimality of the index policy. It is perhaps the most crucial feature of the bandit model used in the proof of the optimality of the index policy. This assumption ensures that an arm, if not played at the current time, promises the same future rewards except for the exogenous discounting. Its reward process is simply frozen in time, waiting to be collected in the future. Arm selection policies thus affect only how the N reward sequences offered by these N arms are interleaved, but not the reward sequences themselves. The best interleaving is the one that mitigates, to the maximum extent, the impact of discounting. It is then intuitive that the optimal policy chooses the arm with the largest reward rate (given by the Gittins index) to ensure that the reward segment with the highest equivalent constant rate is collected first, hence experiencing the least discounting. We further illustrate the necessity of A2.3 with the following counterexample.

CounterExample 3.2 Consider Example 2.4 with $N = 2$ stacks of coins and $\beta = 0.5$. The expected reward sequences offered by these two arms are, respectively, $(0.8, 0, 0, \ldots)$ and $(0.7, 0.6, 0, 0, \ldots)$. We now assume that A2.3 is violated by considering a scenario where each time the second arm is not played, the top coin from this stack is removed without any reward.

It is easy to see that if we follow the Gittins index policy that plays arm 1 at $t = 1$, the reward of 0.7 offered by the top coin in the second stack cannot be realized, and the total discounted reward is $0.8 + 0.6\beta = 1.1$. If instead, we collect the two positive rewards offered by arm 2 first, the total discounted reward would be $0.7 + 0.6\beta + 0.8\beta^2 = 1.2$.

While the Gittins index policy is generally suboptimal when A2.3 is violated, one may question whether there are index forms that take into account the dynamics of passive arms and offer optimality. This question was first explored by Whittle, 1988 [212], when he proposed the *restless* bandit model and developed an index known as the Whittle index. We discuss this important variant of the bandit model in Section 3.3.

3.1.3 MODELING ASSUMPTIONS ON THE REWARD STRUCTURE

A3.1 is necessary. Consider, for example, arm 1 offers an exceptionally high reward when arm 2 is in a particular state. It is then desirable to first drive arm 2 to this particular state even though its current state has a small index. In general, any dependency across arms, be it in the state dynamics or the reward structure, threatens the optimality of an index policy that fundamentally decouples the arms.

A3.2 is not necessary. When passive arms also offer rewards, a simple modification to the Gittins index preserves the index theorem as discussed in Section 3.4. This variant often arises in queueing networks where unserved queues incur holding costs (negative rewards).

A3.3 is necessary. We show in Section 3.4 that a bandit with switching cost can be cast as a special restless bandit problem.

3.1.4 MODELING ASSUMPTIONS ON THE PERFORMANCE MEASURE

The Gittins index policy is optimal in maximizing the (geometrically) discounted reward over an infinite horizon. A change in the performance measure may render it suboptimal. In particular, consider a general discount sequence $\{\alpha_t\}_{t \geq 0}$ and the objective of maximizing the total reward discounted according to $\{\alpha_t\}_{t \geq 1}$:

$$\mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \alpha_t r(S(t), \pi(S(t))) \mid S(1) = s \right]. \tag{3.1}$$

It has been shown by Berry and Fristedt, 1985 [33], that for a general index theorem to hold, the discount sequence must be geometric[1] in the form of $\alpha_t = \beta^{t-1}$ for some $\beta \in (0, 1)$ (see also Section 3.4.2 of Gittins, Glazebrook, and Weber, 2011 [90]).

[1] For continuous-time problems, the discount function must be exponential in the form of $\alpha_t = e^{-\gamma t}$ for some $\gamma > 0$.

This result also excludes the possibility of a general index theorem for the finite horizon problem, which corresponds to a discount sequence of $\alpha_t = 1$ for $t = 1, 2, \dots, T$ and $\alpha_t = 0$ for $t > T$ (see (2.2)). A finite-horizon MDP problem generally dictates a nonstationary policy for optimality: as time approaches the end of the horizon, the optimal decision tends to be more myopic and less forward planning. Being a stationary policy, the suboptimality of the index policy is expected, except in special cases. In particular, the luxury of achieving the maximum equivalent constant reward rate (i.e., the index) by activating an arm until the optimal stopping time τ^* may no longer be feasible if τ^* exceeds the remaining horizon. The optimal action needs to strike a balance between the reward rate and the required time for achieving that rate. The problem has an interesting connection with a stochastic version of the classical NP-hard problem of knapsack, in which a burglar, with a knapsack of a given weight capacity (corresponding to T), needs to choose from items of different values (in this case, the index values as the maximum reward rates) and weights (the stopping times for achieving the index values). It is clear that the optimal solution requires a joint consideration of all available items; any index policy that decouples the items is suboptimal in general.

There are optimality criteria other than the geometrically discounted reward under which the index theorem holds. In Section 3.5, we discuss two such criteria: the total undiscounted reward over a policy-dependent random horizon and the average reward over an infinite horizon. The former leads to the so-called stochastic shortest path bandit model arising in applications concerned with finding the most rewarding (or the least costly) path to a terminating goal such as finding an object or accomplishing a task.

3.2 VARIATIONS IN THE ACTION SPACE

In this section, we discuss several variants of the bandit model that have a more complex action space.

3.2.1 MULTITASKING: THE BANDIT SUPERPROCESS MODEL

The bandit superprocess model relaxes assumption A1.2 and allows multiple control options for activating an arm. This variant and the restless bandit model discussed in Section 3.3 are arguably the two most significant extensions to the bandit model in terms of their generality and applicability. In particular, all forms of arm dependencies, be it dependencies in state evolution, in reward structure, or in terms of precedence constraints, result in special cases of bandit superprocess models. Specifically, dependent arms can be grouped together as a superprocess with control options consisting of playing each constituent arm of the chosen superprocess.

Definition 3.3 *The Superprocess Model:*
Consider N Markov decision processes referred to as superprocesses, with state space $\mathcal{S}_i$, action space $\mathcal{A}_i$, reward function $\{r_i(s,a)\}_{s \in \mathcal{S}_i; a \in \mathcal{A}_i}$, and state transition probabilities $\{p_i(s,s';a)\}_{s,s' \in \mathcal{S}_i; a \in \mathcal{A}_i}$ for $i = 1, 2, \dots, N$. At each time, two control decisions are made. First,

one of the superprocesses, say i, is chosen for continuation; the rest $N - 1$ superprocesses are frozen and offer no reward. Second, a control $a \in \mathcal{A}_i$ is selected and applied to superprocess i. The objective is to maximize the expected total discounted reward with a discount factor β. It is assumed that the control set $\mathcal{A}_i$ is finite for all i.

A juxtaposition of the superprocess model with the bandit model would illuminate the significantly increased complexity of the former. An N-armed bandit problem consists of N Markov *reward* processes, one associated with each arm. The objective is to interleave these N reward processes into a single stream that has a maximum discounted sum. A superprocess model consists of N Markov *decision* processes. It is a *composite* decision problem that controls, at the inner level, the evolution of each of these N MDPs, and at the outer level, the interleave of these constituent decision processes.

To see the connection between the bandit and the superprocess models, consider that the inner-level control is fixed with a specific (stationary deterministic Markov) policy π_i applied to superprocess i for $i = 1, \ldots, N$. The superprocess problem then reduces to an N-armed bandit problem, with each arm given by a Markov reward process characterized by $\{r_i(s, \pi_i(s))\}_{s \in \mathcal{S}_i}$ and $\{p_i(s, s'; \pi_i(s))\}_{s, s' \in \mathcal{S}_i}$. This leads to another way of viewing the two levels of control in the superprocess model. Each superprocess consists of a collection of Markov reward processes, each corresponding to a stationary deterministic Markov policy. The inner control determines which N reward processes, one from each superprocess, to interleave; the outer-level control determines how to interleave. It is perhaps not surprising that in most cases, the optimal composition of the N reward processes can only be determined by considering all N superprocesses jointly. In other words, which reward process (equivalently, which control policy π_i) of a superprocess contributes the most to the discounted sum of the final interleaved single reward stream depends on the other $N - 1$ reward processes that it will be interleaved with. This in general renders an index-type policy suboptimal.

We give below a simple example showing that even when a superprocess is to be interleaved with a reward sequence of a constant value λ, which reward process of the superprocess should be chosen depends on the value of λ.

Example 3.4 Consider a bandit superprocess model consisting of a superprocess and a standard arm offering a constant reward λ. The superprocess has three states: $\mathcal{S} = \{1, 2, 3\}$. When in state 1, there are two control alternatives: a_1 and a_2. Control a_1 changes the state from 1 to 2 and offers a reward of $r(1, a_1)$. Control a_2 changes the state from 1 to 3 and offers a reward of $r(1, a_2)$. States 2 and 3 are absorbing states (i.e., with a self-loop of probability 1) with no control alternatives and offering rewards $r(2)$ and $r(3)$, respectively. Suppose that $r(1, a_2) > r(1, a_1) > r(2) \gg r(3)$ and the initial state is $S(1) = 1$.

It is easy to see that the superprocess with initial state 1 consists of two reward processes $\{r(1, a_1), r(2), r(2), \ldots\}$ and $\{r(1, a_2), r(3), r(3), \ldots\}$, corresponding to applying a_1 and a_2 at state 1, respectively. When the constant reward λ of the standard arm satisfies $\lambda < r(3)$, the first reward sequence of the superprocess should be selected (i.e., the optimal control at state 1 is a_1) and the standard arm will never be activated. This leads to a total discounted reward of $r(1, a_1) + \beta \frac{r(2)}{1-\beta}$. When $r(1, a_2) > \lambda > r(1, a_1)$, it is optimal to select the second reward sequence of the superprocess by applying a_2 at the initial state 1 and then switching to the standard arm. This yields a total discounted reward of $r(1, a_2) + \beta \frac{\lambda}{1-\beta}$.

When multiple loosely coupled decision problems competing for early attention (thus less discounting), the optimal strategy tends to be more myopic than the optimal policies for each individual decision problems. Adopting the optimal policy for each decision problem, and then using the index policy as the outer-level control for allocation, is in general suboptimal.

Nevertheless, the hope for an index-type optimal policy is not all lost. Next, we show a necessary and sufficient condition for the index theorem and give a natural extension of the Gittins index to a superprocess. We start with the latter.

For each of the Markov reward process associated with a given control policy π_i for superprocess i with initial state s, we can obtain its Gittins index as given in (2.20):

$$\nu_i(s; \pi_i) = \max_{\tau \geq 1} \frac{\mathbb{E}_{\pi_i}\left[\sum_{t=1}^{\tau} \beta^{t-1} r_i(S_i(t), \pi_i(S_i(t))) \;\middle|\; S_i(1) = s\right]}{\mathbb{E}_{\pi_i}\left[\sum_{t=1}^{\tau} \beta^{t-1} \;\middle|\; S_i(1) = s\right]}. \tag{3.2}$$

A natural candidate for the index of superprocess i at state s is thus

$$\nu_i(s) = \max_{\pi_i} \nu_i(s; \pi_i), \tag{3.3}$$

the largest Gittins index among all reward processes of superprocess i. Let $a_i^*(s)$ denote the optimal control at state s that attains $\nu_i(s)$. The index policy is thus as follows: given the current states $(s_1, s_2, \ldots, s_N)$ of the N superprocesses, the process i^* with the greatest index value $\nu_{i^*}(s_{i^*}) = \max_i \nu_i(s_i)$ is activated with control $a_{i^*}^*(s_{i^*})$.

In order for any index policy, which decouples the superprocesses, to be optimal, it is necessary that whenever this index policy activates superprocess i at state s, it chooses the same control a, irrespective of the states of other superprocesses. In other words, each superprocess has a *dominating* reward process in its collection of Markov reward processes that should be chosen regardless of the $N-1$ reward processes that it will be interleaved with. It was shown by Whittle, 1980 [210] that this condition holds if and only if the same can be said about superprocess i when it is to be interleaved with a constant reward sequence. We state this condition below.

Condition 3.5 *Whittle Condition for the Superprocess Model:*
A superprocess is said to satisfy the Whittle Condition if in a bandit superprocess model consisting of this superprocess and a standard arm offering a constant reward λ, the optimal control policy for the superprocess is independent of λ. More specifically, there exists a control policy π^* (a mapping from the state space $\mathcal{S}$ of the superprocess to the control set $\mathcal{A}$) such that, for all s and λ for which it is optimal to activate the superprocess, it is optimal to apply control $a = \pi^*(s)$.

When the Whittle condition holds for each constituent superprocess, the problem reduces to a bandit problem by considering, without loss of optimality, only the dominating reward process of each superprocess generated by its dominating policy π_i^* (as defined in the Whittle condition). The policy associated with the index

$$\nu_i(s) = \max_{\pi_i} \nu_i(s; \pi_i) = \nu_i(s; \pi_i^*) \tag{3.4}$$

is thus optimal.

Theorem 3.6 *Consider the superprocess model defined in Definition 3.3. Suppose that every superprocess satisfies the Whittle condition. The index policy with respect to the index defined in* (3.3) *is optimal.*

This condition, however, is a strong condition. Example 3.4 gives a glimpse into how easily it can be violated. There are, however, special classes of the superprocess model with rather general applicability that satisfies the Whittle Condition. For example, the class of stoppable bandits introduced by Glazebrook, 1979 [92].

Glazebrook, 1982 [93], showed that a general bandit superprocess model does not admit an index-type optimal policy. An upper bound on the value function of a general superprocess problem was established by Brown and Smith, 2013 [44], based on the so-called *Whittle Integral* constructed by Whittle, 1980 [210]. Hadfield-Menell and Russell, 2015 [96], developed efficient algorithms for computing the upper bound and ϵ-optimal policies.

3.2.2 BANDITS WITH PRECEDENCE CONSTRAINTS

We now discuss relaxing modeling assumption A1.3 by allowing precedence constraints that restrict the availability of arms. Bandits with precedence constraints often arise in stochastic scheduling of jobs/projects. Jobs are the constituent arms in the resulting bandit model. Precedence constraints, however, lead to dependencies among arms, and the problem falls within the superprocess model. Being a special class of the superprocess model, however, it preserves the index theorem when the precedence constraints or the job characteristics satisfy certain conditions.

Jobs, precedence constraints, and the superprocess formulation: A generalized job is a bandit process with a terminating state representing the completion of the job beyond which no reward may be accrued. A job that yields a single reward at the time of completion and no rewards before or after completion is called a simple job. A simple job is thus characterized by a random processing time X and a reward r earned upon completion.

For a collection of N jobs denoted by $\{1, 2, \ldots, N\}$, a set $\mathcal{C}$ of precedence constraints has elements of the form $\{i > j\}$ representing the restriction that the service of job j can only commence after the completion of job i. Only scheduling strategies that comply with $\mathcal{C}$ are admissible.

A precedence constraint set $\mathcal{C}$ can be represented by a digraph $D_{\mathcal{C}}$. The vertices of the digraph are the jobs. An arc (i.e., a directed edge) exists from vertex i to vertex j if and only if $\mathcal{C}$ contains the precedence constraint $\{i > j\}$. Figures 3.1 and 3.2 show two examples.

Each connected component of $D_{\mathcal{C}}$ constitutes a superprocess. The state of the superprocess is the Cartesian product of the states of the jobs in this connected component. The control options are in choosing which unfinished job to serve when this superprocess is activated. In the example shown in Figure 3.2, there are two superprocesses, each consisting of jobs $\{1, \ldots, 6\}$ and $\{7, \ldots, 11\}$, respectively.

Scheduling generalized jobs with preemption: We start with the most general formulation of stochastic scheduling: jobs are general bandit processes with the addition of terminating states, the precedence constraints are arbitrary, and preemption is allowed (i.e., an unfinished job can

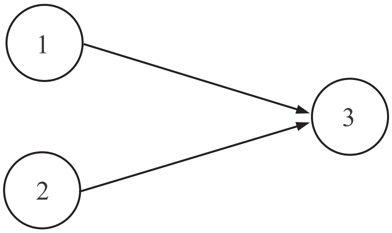


Figure 3.1: Precedence constraints invalidating the index theorem.

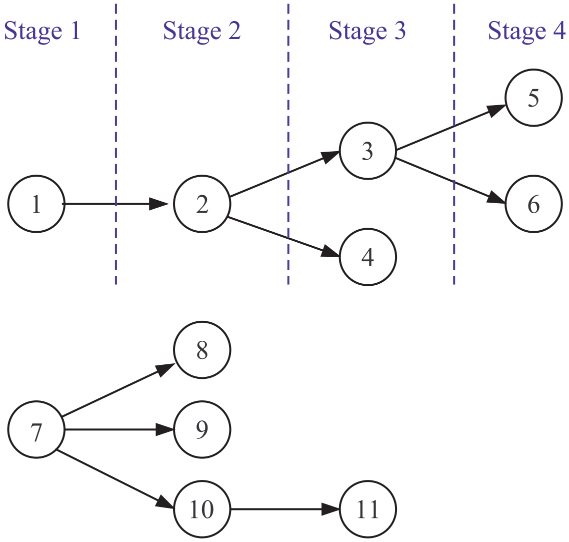


Figure 3.2: Precedence constraints forming an out forest.

be interrupted and resumed at a later time without any loss). For this general problem, the index theorem does not hold, as shown by the following counter example constructed based on Example 4.1 from Gittins, Glazebrook, and Weber, 2011 [90].

CounterExample 3.7 Consider three jobs $\{1, 2, 3\}$ with required service time X_i and reward r_i upon completion ($i = 1, 2, 3$). Suppose that X_1 takes values of 1 and M ($M \gg 1$) with equal probability, $X_2 = X_3 = 1$, and $r_1 = 1$, $r_2 = 2$, and $r_3 = 64$. The precedence constraints are such that job 3 can only start after jobs 1 and 2 are completed (see Figure 3.1). Rewards are discounted such that completion of job i at time t contributes $\beta^{t-1} r_i$ to the total payoff, where the discount factor $\beta = \frac{1}{2}$. Preemption is allowed.

The problem can be modeled as a single superprocess consisting of all three jobs. Each scheduling policy π results in a Markov reward process with associated Gittins indices as given in (3.2). The scheduling strategy chosen under the index rule is the policy that maximizes the index in the given state (see (3.3)). We show next that the scheduling policy determined this way is suboptimal.

Since $r_2 > r_1$, it is easy to see that the optimal strategy π^* is to schedule job 2 first, followed by job 1 and then job 3 upon the completion of job 1. The index policy, however, would schedule job 1 first as detailed below.

Let s_0 denote the initial state of this superprocess at $t = 1$, when no job has received any service time. To compute the index $\nu(s_0)$, it suffices to compare two policies: policy π^* specified above and policy π' that schedules job 1 at $t = 1$, switches to job 2 at $t = 2$, and then switch back to job 1 at $t = 2$ if it has not been completed, and then job 3. All other admissible policies would yield index values smaller than that of π^* or π'. It can be shown that the optimal stopping times attaining the indices $\nu(s_0; \pi^*)$ and $\nu(s_0; \pi')$ are at the time when job 1 receives 1 unit of service time if it fails to terminate, and otherwise are at the completion of all jobs (i.e., $t = 3$). To simplify the calculation of the indices, we assume $M = \infty$, which does not affect the ordering of the indices for sufficiently large M. We thus have

$$\nu(s_0; \pi^*) = \frac{\mathbb{P}[X_1 = 1]\left(r_2 + \beta r_1 + \beta^2 r_3\right) + \mathbb{P}[X_1 = M] r_2}{\mathbb{P}[X_1 = 1](1 + \beta + \beta^2) + \mathbb{P}[X_1 = M](1 + \beta)} \simeq 6.31 \tag{3.5}$$

$$\nu(s_0; \pi') = \frac{\mathbb{P}[X_1 = 1]\left(r_1 + \beta r_2 + \beta^2 r_3\right) + \mathbb{P}[X_1 = M] \times 0}{\mathbb{P}[X_1 = 1](1 + \beta + \beta^2) + \mathbb{P}[X_1 = M] \times 1} \simeq 6.55. \tag{3.6}$$

Since $\nu(s_0; \pi') > \nu(s_0; \pi^*)$, the index policy thus follows π'.

The reason behind the suboptimality of the index policy in the above example is the dependency, through the common out-neighbor (i.e., job 3 in the precedence digraph depicted in Figure 3.1), between the reward streams associated with scheduling job 1 and scheduling job 2 at $t = 1$. These two alternatives thus cannot be considered in isolation by comparing their indices for choosing the optimal action.

A natural question is thus what topological structures of the precedence digraph would preserve the index theorem. CounterExample 3.7 shows that the key is to ensure that at any given time, all jobs ready for service should not share common out-neighbors. This is satisfied by precedence constraints that form an *out-forest* as defined below.

Definition 3.8 An out-forest is a directed graph with no directed cycles and in which the in-degree of each vertex is at most one.

The directed graph in Figure 3.2 is an out-forest with two out-trees. Since the in-degree of each vertex is at most one, no two vertices in an out-forest share a common out-neighbor.

The index policy for scheduling generalized jobs with preemption and with precedence constraints forming an out-forest can be obtained through an interesting combination of backward induction with forward induction. We describe the policy using the example in Figure 3.2.

The two connected components in Figure 3.2 constitute the two superprocesses. The first step is to reduce each superprocess to a single Markov reward process, hence a bandit process, by fixing the inner-level control. Take the first superprocess as an example. The precedence constraints partition the decision horizon into four stages, punctuated in sequence by the completion times of jobs 1, 2, and 3. The optimal policy for this MDP can be obtained by combining a backward induction over these four stages and a forward induction (i.e., the Gittins index policy) within each stage. Specifically, the backward induction starts with the fourth stage, which begins upon the completion of job 3. The remaining decision problem is the scheduling of two sink vertices—jobs 5 and 6—that are two independent reward processes. The problem is thus a two-armed bandit for which the index policy is optimal. Let $5 \oplus 6$ denote the reward process resulting from interleaving the reward streams of job 5 and job 6 based on the Gittins index policy. We now move backward in time to the beginning of the third stage at the completion time of job 2. The decision problem starting from this time instant is the scheduling of jobs 3 and 4. The problem is again a two-armed bandit, in which the reward process associated with scheduling job 3 is $(3,\ 5 \oplus 6)$ that concatenates the reward process of job 3 with $5 \oplus 6$ to account for the entire remaining horizon. The optimal strategy is the index policy that results in a single reward process $(3,\ 5 \oplus 6) \oplus 4$ that optimally interleaves the reward process of $(3,\ 5 \oplus 6)$ and that of job 4. Continue the backward induction to $t = 1$. We have reduced this superprocess to a single reward process given by $(1,\ 2,\ ((3,\ 5 \oplus 6) \oplus 4)$. Repeat this procedure for the second superprocess consisting of jobs $\{7, \ldots, 11\}$. The outer-level control for scheduling these two superprocesses is thus reduced to a two-armed bandit problem with the two reward processes obtained in the first step.

The following theorem, due to Glazebrook, 1976 [91], states the optimality of the index policy.

Theorem 3.9 *For the superprocess model for scheduling generalized jobs with preemption and with precedence constraints forming an out-forest, the index policy is optimal.*

3.2.3 OPEN BANDIT PROCESSES

The canonical bandit model is closed in the sense that the set of arms is fixed over the entire decision horizon. Nash, 1973 [153], Whittle, 1981 [211], and Weiss, 1988 [209] considered open bandit processes where new arms are continually appearing. Such open bandit processes are referred to as arm-acquiring bandits by Whittle, 1981 [211], and branching bandit processes by Weiss, 1988 [209].

Whittle provided a complete extension of the index theorem to open bandit processes under the assumption that the arrival processes of new arms are i.i.d. over time and independent of the states of currently present arms, all past actions, and the state transition of the activated arm. Weiss considered a more general model where the arrivals may depend on the state of the arm currently being activated.

It is worth pointing out that while new arms continually appearing and the total number of arms growing indefinitely, the number of arm *types* as characterized by state transition probabilities and reward functions is fixed and finite. The open process also complicates the characterization of the index form. Differing from the closed model, the index associated with an arm type involves new arms of different types that may enter the system. Lai and Ying, 1988 [127] showed that under certain stability assumptions, the index rule of a much simpler closed bandit process is asymptotically optimal for the open bandit process as the discount factor β approaches 1.

3.3 VARIATIONS IN THE SYSTEM DYNAMICS

As discussed earlier, modeling assumption A2.1 on the Markovian dynamics of active arms is not necessary for the index theorem. A2.3 is, however, crucial to the optimality of Gittins index. The next natural question to ask is whether a different index form might offer optimality when A2.3 does not hold. This question was first considered by Whittle, 1988 [212], who proposed a *restless* bandit model that allows passive arms to evolve. This variant is the focus of this section.

3.3.1 THE RESTLESS BANDIT MODEL

Definition 3.10 *The Restless Multi-Armed Bandit Model:*
Consider N arms, each with state space $\mathcal{S}_i$ $(i = 1, \ldots, N)$. At time $t = 1, 2, \ldots$, based on the observed states $[S_1(t), \ldots, S_N(t)]$ of all arms, K $(1 \leq K < N)$ arms are activated. The reward and the state transition probabilities of arm i at state s are $\{r_i(s),\ p_i(s, s')\}_{s,s' \in \mathcal{S}_i}$ when it is active and $\{\mathring{r}_i\ (s),\ \mathring{p}_i\ (s, s')\}_{s,s' \in \mathcal{S}_i}$ when it is passive. The performance measure can be set to either the expected total discounted reward with a discount factor β $(0 < \beta < 1)$ or the average reward, both over an infinite horizon.

The above restless bandit model extends the original bandit model in three aspects: simultaneous activation of multiple arms, rewards offered by passive arms, and state evolution of passive arms. Allowing passive arms to change state is the most salient feature of the restless bandit model, hence the name given. Simultaneous play of K arms, while alone would render Gittins index policy suboptimal, is not essential to the general theory for restless bandits. The general results apply equally well for $K = 1$ and $K > 1$, with the exception that the asymptotic optimality of the Whittle index policy requires K growing proportionally with N as N approaching infinity. Allowing passive arms to offer rewards does not change the problem in any essential way; one may, without loss of generality, assume $\overset{\circ}{r}_i\ (s) = 0$ for all i and s. As we will see in Section 3.4, a simple modification to the Gittins index preserves the index theorem for a bandit model with rewards under passivity. A similar argument holds for the restless bandit model.

The restless bandit model significantly broadens the applicability of the bandit model in diverse application domains. For instance, in communication networks, the condition of a communication channel or routing path may change even when it is not utilized. In queueing systems, queues continue to grow due to new arrivals even when they are not served. In target tracking, targets continue to move when they are not monitored. These applications fall under the restless bandit model.

3.3.2 INDEXABILITY AND WHITTLE INDEX

To introduce the concepts of indexability and the Whittle index, it suffices to consider a single arm. We thus omit the subscript for the arm labels from all notations. We consider the discounted reward criterion. The treatment of the average reward case can be similarly obtained.

Definition 3.11 *A Single-Armed Restless Bandit Problem:*
Consider a single-armed restless bandit process with reward functions and state transition probabilities given by $\{r(s),\ p(s, s')\}_{s,s' \in \mathcal{S}}$ when it is active and $\{\overset{\circ}{r}\ (s),\ \overset{\circ}{p}\ (s, s')\}_{s,s' \in \mathcal{S}}$ when it is passive. The objective is an activation policy that maximizes the expected total discounted reward with a discount factor β $(0 < \beta < 1)$.

A single-armed restless bandit process is an MDP with two possible actions—$\{0$ (passive), 1 (active)$\}$—corresponding to whether the arm is made active or passive. The optimal policy π^*, which is a stationary deterministic Markov policy (see Theorem 2.1), defines a partition of the state space $\mathcal{S}$ into a passive set $\{s : \pi^*(s) = 0\}$ and an active set $\{s : \pi^*(s) = 1\}$, where $\pi^*(s)$ denotes the action at state s under policy π^*.

Whittle index measures how rewarding it is to activate the arm at state s based on the concept of *subsidy for passivity*. Specifically, we modify the single-armed restless bandit process by introducing a subsidy $\lambda \in \mathbb{R}$ which is a constant additional reward accrued each time the arm is made passive. This subsidy λ changes the optimal partition of the passive and active sets: when λ approaches negative infinity, the optimal passive set approaches an empty set; when λ approaches positive infinity, the optimal passive set approaches the entire state space $\mathcal{S}$. Intuitively, a state that requires a greater subsidy λ (i.e., a greater additional incentive) for it to join the passive set should enjoy higher priority for activation. This is the basic idea behind the Whittle index which we define below.

Let $V_\lambda(s)$ denote the value function representing the maximum expected total discounted reward that can be accrued from the single-arm restless bandit process with subsidy λ when the initial state is s. Considering the two possible actions at the first decision time, we have

$$V_\lambda(s) = \max\{Q_\lambda(s; a = 0),\ Q_\lambda(s; a = 1)\}, \tag{3.7}$$

where $Q_\lambda(s; a)$ denotes the expected total discounted reward obtained by taking action a at $t = 1$ followed by the optimal policy for $t > 1$ and is given by

$$\begin{aligned} Q_\lambda(s; a = 0) &= (\overset{\circ}{r}(s) + \lambda) + \beta \sum_{s' \in \mathcal{S}} \overset{\circ}{p}(s, s') V_\lambda(s'), && (3.8) \\ Q_\lambda(s; a = 1) &= r(s) + \beta \sum_{s' \in \mathcal{S}} p(s, s') V_\lambda(s'). && (3.9) \end{aligned}$$

The first terms in (3.8) and (3.9) are the immediate rewards obtained at $t = 1$ under the respective actions of $a = 0$ and $a = 1$. The second terms are the total discounted rewards in the remaining horizon determined by the value function due to the adoption of the optimal policy for $t > 1$. The optimal policy π^*_λ under subsidy λ is given by

$$\pi^*_\lambda(s) = \begin{cases} 1, & \text{if } Q_\lambda(s; a = 1) > Q_\lambda(s; a = 0) \\ 0, & \text{otherwise} \end{cases}, \tag{3.10}$$

where we have chosen to make the arm passive when the two actions are equally rewarding. The passive set $\mathcal{P}(\lambda)$ under subsidy λ is given by

$$\mathcal{P}(\lambda) = \{s :\ \pi^*_\lambda(s) = 0\} = \{s :\ Q_\lambda(s; a = 0) \geq Q_\lambda(s; a = 1)\}. \tag{3.11}$$

It is intuitive to expect that as the subsidy λ increases, the passive set $\mathcal{P}(\lambda)$ grows monotonically. In other words, if the arm is made passive at state s under a subsidy λ, it should also be made passive at this state under a subsidy $\lambda' > \lambda$. Unfortunately, as shown in the counterexample below, this need not hold in general.

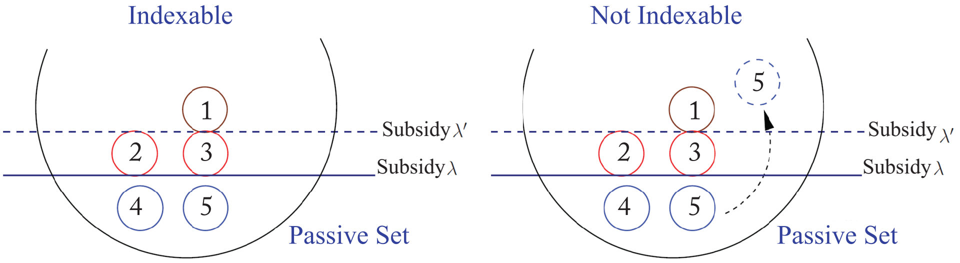
CounterExample 3.12 Consider an arm with state space $\{1, 2, 3\}$. At $s = 1$, the state transits, deterministically, to 2 and 3 under $a = 1$ and $a = 0$, respectively; states 2 and 3 are absorbing states regardless of the actions. The rewards are given by $r(1) = \overset{\circ}{r}(1) = 0$, $r(2) = 0$, $\overset{\circ}{r}(2) = 10$, $r(3) = 11$, $\overset{\circ}{r}(3) = 0$. The discount factor is $\beta = \frac{1}{2}$.

It is not difficult to see that when the subsidy λ is set to 0, the optimal action at $s = 1$ is to be passive so that the arm enters state 3 and we collect the highest reward $r(3) = 11$ by choosing the active action at every subsequent time. When the subsidy increases to $\lambda = 2$, however, the optimal action at $s = 1$ is to be *active* so that the arm enters state 2 and we collect the reward plus subsidy $\overset{\circ}{r}(2) + \lambda = 12$ at every subsequent time. The passive set $\mathcal{P}(\lambda)$ actually *decreases* from $\{1, 2\}$ to $\{2\}$ when the subsidy increases from $\lambda = 0$ to $\lambda = 2$.

When the passive set $\mathcal{P}(\lambda)$ is not monotonic in λ, the subsidy for passivity does not induce a consistent ordering of the states in terms of how rewarding it is to activate the arm, thus may not be a meaningful measure for the index. This is the rationale behind the notion of *indexability*.

Definition 3.13 *Indexability of Restless Bandit Model:*
An arm in a restless bandit model is *indexable* if the passive set $\mathcal{P}(\lambda)$ under subsidy λ increases monotonically from $\emptyset$ to $\mathcal{S}$ as λ increases from $-\infty$ to $+\infty$. A restless multi-armed bandit model is indexable if every arm is indexable.

We illustrate the concept of indexability in Figure 3.3. When an arm is indexable, there is a consistent ordering of all its states in terms of when each state enters the passive set $\mathcal{P}(\lambda)$ as the subsidy λ increases from $-\infty$ to $+\infty$. As illustrated in Figure 3.3, imagine the states as pebbles in an urn. If we liken increasing the subsidy λ to pouring water into the urn, the order at which the pebbles (states) get submerged in water (enter the passive set) induces a consistent and meaningful measure on the positions of the pebbles above the bottom of the urn (how rewarding it is to be active at each state). If an arm is not indexable, then there exists a state such as state 5 in Figure 3.3 that jumps out of "water" (the passive set) as the water level (the subsidy) increases from λ to λ'. The value λ at which each state enters the passive set does not induce a consistent ordering of the states. If it does not provide a consistent measure for comparing states, it may not be a valid candidate for the index that aims to rank arms based on their states (consider, for example, all the arms are stochastically identical with the same state space).


Indexable

Not Indexable

Figure 3.3: Indexability and monotonicity of the passive set of a restless bandit model.

If an arm is indexable, its *Whittle index* $\nu(s)$ of state s is defined as the least value of the subsidy λ under which it is optimal to make the arm passive at s, or equivalently, the minimum subsidy λ that makes the passive and active actions equally rewarding:

$$\nu(s) = \inf_{\lambda} \left\{\lambda : \pi_{\lambda}^{*}(s) = 0\right\} = \inf_{\lambda} \{\lambda : Q_{\lambda}(s; a = 0) = Q_{\lambda}(s; a = 1)\}. \tag{3.12}$$

The definitions of indexability and Whittle index direct lead to the following result on the optimal policy for a single-armed restless bandit problem.

Theorem 3.14 *For the single-armed restless bandit problem given in Definition 3.11, if it is indexable, then the policy that activates the arm if and only if the current state has a positive Whittle index is optimal, i.e.,*

$$\pi^*(s) = \begin{cases} 1 & \textit{if } \nu(s) > 0 \\ 0 & \textit{if } \nu(s) \leq 0 \end{cases} . \tag{3.13}$$

Similarly, with a given subsidy λ, activating the arm at states whose Whittle indices exceed λ is optimal.

A discussion on the relation between the Whittle index and the Gittins index is in order. The bandit model is a special case of the restless bandit model with $\overset{\circ}{p}(s,s) = 1$ and $\overset{\circ}{r}(s) = 0$ for all s and for all arms. When a restless bandit problem reduces to a bandit problem, the Whittle index reduces to the Gittins index. This can be seen by noticing that once the passive action is taken at a particular time t_0, it will be taken at every $t > t_0$ since the state of the arm is frozen. Thus, a constant subsidy λ is to be collected at each time $t > t_0$, which can be viewed as switching to a standard arm offering a constant reward λ. The Whittle index induced by the subsidy for passivity thus reduces to the Gittins index calibrated by a standard arm with a constant reward.

A less obvious fact is that the canonical bandit model defined in Definition 2.8 is indexable under Whittle's definition given in Definition 3.13. This was established by Whittle in his original paper on restless bandits (Whittle, 1988 [212]). The basic idea is to show that a state s in the passive set remains in the passive set as the subsidy increases. This is equivalent to a positive slope of the gain under $a = 0$ over $a = 1$ for $\lambda > \nu(s)$, i.e.,

$$\frac{\partial}{\partial \lambda}\left(Q_\lambda(s; a = 0) - Q_\lambda(s; a = 1)\right) \geq 0 \tag{3.14}$$

for $\lambda > \nu(s)$. The above can be shown based on (3.8)–(3.9) and the fact that the derivative of the value function with respect to λ is the expected total discounted time that the arm is made passive. The latter is quite intuitive: when the subsidy for passivity λ increases, the rate at which the total discounted reward $V_\lambda(s)$ increases is determined by how often the arm is made passive. The proof completes by noticing that the total discounted time in passivity starting at a state s in the passive set (i.e., $\lambda > \nu(s)$) is the entire discounted horizon $\frac{1}{1-\beta}$ due to the frozen state under passivity in the bandit model.

For a general restless bandit problem, establishing its indexability can be complex. General sufficient conditions remain elusive. Nevertheless, a number of special classes of the restless bandit model have been shown to be indexable. See, for example, Glazebrook, Ruiz-Hernandez, and Kirkbride, 2006 [94], for specific restless bandit processes that are indexable and Nino-Mora, 2001, 2007b [154, 158], for a numerical approach for testing indexability and calculating Whittle index for finite-state bandit problems. In Section 6.1.1, we discuss a restless bandit model arising in communication networks for which the indexability can be established analytically and the Whittle index obtained in closed form.

3.3.3 OPTIMALITY OF WHITTLE INDEX POLICY

We now return to the general restless bandit model with $N > 1$ arms defined in Definition 3.10. The Whittle index policy for a general restless bandit problem is to activate the K arms of currently greatest Whittle indices. It is in general suboptimal. In this section, we discuss several scenarios in which its optimality can be established.

Optimality under a relaxed constraint: In the original restless bandit model, a hard constraint on the number of active arms is imposed for each decision time. Consider a relaxed scenario where the constraint is imposed on the expected total number of arms activated over the entire horizon rather than at each given time. Specifically, suppose that a unit cost is incurred each time an arm is activated. The constraint is on the total cost, discounted in the same way as the rewards:

$$\mathbb{E}_\pi\left[\sum_{i=1}^{N}\sum_{t=1}^{\infty}\beta^{t-1}\mathbb{I}(a_i(t)=1)\,|\,\mathbf{S}(1)=\mathbf{s}\right]=\frac{K}{1-\beta}, \tag{3.15}$$

where $\mathbf{S}(1) = [S_1(1), \ldots, S_N(1)]$ is the state vector. Activating exactly K arms each time is one way to satisfy this constraint. We thus have the following relaxed MDP problem that admits a larger set of policies:

$$\begin{array}{ll}\text{maximize} & \mathbb{E}_\pi\left[\sum_{i=1}^{N}\sum_{t=1}^{\infty}\beta^{t-1}r_i(S_i(t);a_i(t))\,|\,\mathbf{S}(1)=\mathbf{s}\right]\\ \text{subject to} & \mathbb{E}_\pi\left[\sum_{i=1}^{N}\sum_{t=1}^{\infty}\beta^{t-1}\mathbb{I}(a_i(t)=0)\,|\,\mathbf{S}(1)=\mathbf{s}\right]=\frac{N-K}{1-\beta},\end{array} \tag{3.16}$$

where we have changed the constraint to an equivalent one on the number of passive arms for reasons that will become clear later. The theorem below gives the optimal policy for this relaxed problem in terms of the Whittle index.

Theorem 3.15 *Let $\bar{V}(\mathbf{s})$ denote the value function of the relaxed restless bandit problem given in (3.16). We have*

$$\bar{V}(\mathbf{s})=\inf_{\lambda}\left\{\sum_{i=1}^{N}V_\lambda^{(i)}(s_i)-\lambda\frac{N-K}{1-\beta}\right\}, \tag{3.17}$$

where $V_\lambda^{(i)}(s_i)$ is the value function of arm i with subsidy λ as given in (3.7). This optimal value is achieved by a policy that activates at each time all arms whose current Whittle indices exceed a threshold λ^*. This threshold λ^* is the λ value that achieves the minimum in (3.17), which ensures that the constraint in (3.16) is met.

Proof. The proof is based on a Lagrangian relaxation of the constrained optimization given in (3.16). The resulting Lagrange function with multiplier λ is

$$\mathbb{E}_\pi\left[\sum_{i=1}^{N}\sum_{t=1}^{\infty}\beta^{t-1}r_i(S_i(t);a_i(t))\,|\,\mathbf{S}(1)=\mathbf{s}\right] \tag{3.18}$$

$$+\lambda\left(\mathbb{E}_\pi\left[\sum_{i=1}^{N}\sum_{t=1}^{\infty}\beta^{t-1}\mathbb{I}(a_i(t)=0)\,|\,\mathbf{S}(1)=\mathbf{s}\right]-\frac{N-K}{1-\beta}\right) \tag{3.19}$$

$$= \quad \mathbb{E}_\pi\left[\sum_{i=1}^{N}\sum_{t=1}^{\infty}\beta^{t-1}(r_i(S_i(t);a_i(t))+\lambda\mathbb{I}(a_i(t)=0))\,|\,\mathbf{S}(1)=\mathbf{s}\right]-\lambda\frac{N-K}{1-\beta} \tag{3.20}$$

$$= \quad \sum_{i=1}^{N}V_{\pi,\lambda}^{(i)}(s_i)-\lambda\frac{N-K}{1-\beta}, \tag{3.21}$$

where $V_{\pi,\lambda}^{(i)}(s_i)$ denote the expected total discounted reward obtained from arm i under policy π when a subsidy λ is obtained whenever the arm is made passive. Note that in the above Lagrange function, arms are decoupled except sharing a common value λ of the subsidy. The optimal policy is thus the one that maximizes the discounted reward obtained from each subsidized arm. From Theorem 3.14, the optimal policy is to activate arm i $(i=1,\ldots,N)$ if and only if its current Whittle index exceeds the subsidy λ^* which is the critical value of the Lagrangian multiplier. The expression for λ^* follows from the fact that $V_\lambda^{(i)}(s_i)$, being the maximum of linear functions $V_{\pi,\lambda}^{(i)}(s_i)$ of λ, is convex and increasing in λ. □

The above proof shows clearly the interpretation of the subsidy λ as the Lagrange multiplier for the relaxed constraint given in (3.16). Under the relaxed constraint, the Whittle index policy is optimal.

Let $V(\mathbf{s})$ and $V_W(\mathbf{s})$ denote, respectively, the value of the optimal policy and that of the Whittle index policy for the original restless bandit model with the strict activation constraint. It is easy to see that

$$V_W(\mathbf{s})\leq V(\mathbf{s})\leq\bar{V}(\mathbf{s}), \tag{3.22}$$

i.e., $\bar{V}(\mathbf{s})$ as specified in (3.17) provides an upper bound on the optimal performance. It can be used as a benchmark for gauging the performance of policies under the original restless bandit model.

Asymptotic optimality: Under the strict constraint that exactly K arms be activated each time, Whittle conjectured that the index policy is asymptotically optimal under the average reward criterion. Suppose that the N arms are of L different types. Arms of the same type l are stochastically identical with the same Markovian dynamics $\{p_l(s,s'),\ \overset{\circ}{p}_l\,(s,s')\}$ and the same reward functions $\{r_l(s)\}$ and $\{\overset{\circ}{r}_l\,(s)\}$. Consider that as N approaches infinity, the composition of the arms approaches to a limit $\{\alpha_l\}_{l=1}^{L}$, where α_l is the proportion of type-l arms. For such a population of arms with a limiting composition, we are interested in the average reward per arm as N approaches infinity and K grows proportionally, i.e., K/N tends to a limit κ.

Consider first a single arm of type l. Let $g_l(\lambda)$ denote the optimal average reward offered by this arm with subsidy λ:

$$g_l(\lambda) = \sup_{\pi} \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}_{\pi}$$
$$\left[\sum_{t=1}^{T} r_l(S_t)\mathbb{I}(\pi(S(t)) = 1) + (\lambda + \overset{\circ}{r}_l\,(S(t)))\mathbb{I}(\pi(S(t)) = 0)\,\middle|\, S(1) = s\right], \qquad (3.23)$$

where the subscript l denotes the arm type rather than the arm label. We have also assumed, for simplicity, that the arm states communicate well enough (more specifically, all policies result in a Markov chain with a single recurrent class) that $g_l(\lambda)$ is independent of the initial state. A corresponding result of Theorem 3.14 under the average reward criterion states that $g_l(\lambda)$ is achieved by activating the arm at states with Whittle indices exceeding λ, where the Whittle index is defined under the average reward criterion.

Let $\bar{g}(N,\kappa)$ denote the optimal average reward under the relaxed constraint that the time-averaged number of active arms equals κN. Based on a result similar to Theorem 3.15 for the average reward criterion, we have

$$\lim_{N\to\infty} \frac{\bar{g}(N,\kappa)}{N} = \inf_{\lambda}\left[\sum_{l=1}^{L} \alpha_l g_l(\lambda) - \lambda(1-\kappa)\right]. \qquad (3.24)$$

Let $g_W(N,\kappa)$ denote the average reward per arm obtained by the Whittle index policy that activates exactly the $K = \kappa N$ arms of the greatest indices. We have the following conjecture by Whittle.

Conjecture 3.16 Suppose that all arms are indexable. Then

$$\lim_{N\to\infty} g_W(N,\kappa) = \lim_{N\to\infty} \frac{\bar{g}(N,\kappa)}{N}. \qquad (3.25)$$

Since the relaxed problem defines an upper bound on the original restless bandit model, (3.25) implies the asymptotic optimality of the Whittle index policy in terms of average reward per arm. The rationale behind this conjecture is based on the law of large numbers. For sufficiently large N, the number of arms with indices exceeding the threshold λ^* (the value of λ achieving the infimum in (3.24)) deviates from κ only by a term of order at most $N^{-\frac{1}{2}}$. The performance loss *per arm* of the Whittle index policy as compared to the optimal policy under the relaxed constraint thus diminishes as N approaches infinity.

Based on a fluid approximation of the Markov processes resulting from the Whittle index policy, Weber and Weiss, 1990, 1991 [206, 207], have shown that the conjecture is true if the differential equations of the fluid approximation have an asymptotic stable equilibrium point. This condition can be difficult to check for a general restless bandit model, but is shown to hold for all restless bandits with a state space size no greater than 3.

Optimality in special cases: A number of interesting cases exist in the literature in which the Whittle index policy was shown to be optimal or near optimal. For instance, Lott and Teneketzis, 2000 [139], cast the problem of multichannel allocation in single-hop mobile networks with multiple service classes as a restless bandit problem and established sufficient conditions for the optimality of a myopic-type index policy. Raghunathan et al., 2008 [166], considered multicast scheduling in wireless broadcast systems with strict deadlines and established the indexability and obtained the analytical form of the Whittle index for the resulting restless bandit problem. Ehsan and Liu, 2004 [76], formulated a bandwidth allocation problem in queuing systems as a restless bandit and established the indexability. They further obtained the Whittle index in closed-form and established sufficient conditions for the optimality of the Whittle index policy. The restless bandit model has also been applied to economic systems for handling inventory regulation by Veatch and Wein, 1996 [201]. In Section 6.1.1, we consider a restless bandit problem arising from communication networks for which indexability and the performance of the Whittle index policy can be established analytically.

3.3.4 COMPUTATIONAL APPROACHES TO RESTLESS BANDITS

In a series of papers, Bertsimas and Nino-Mora, 2000 [35] and Nino-Mora, 2001, 2002, 2006 [154–156] developed algorithmic and computational approaches based on linear programming for establishing indexability and computing priority indexes for a general restless bandit model. A comprehensive survey on this line of work can be found in Nino-Mora, 2007 [158]. Glazebrook, Ruiz-Hernandez, and Kirkbride, 2006 [94] gave several examples of indexable families of restless bandit problems.

3.4 VARIATIONS IN THE REWARD STRUCTURE

3.4.1 BANDITS WITH REWARDS UNDER PASSIVITY

The assumption that passive arms generate no reward is not necessary for the optimality of the Gittins index policy. This is welcome news in particular to applications such as queueing systems and job scheduling where a holding cost incurs for each unserved queue/job. This type of bandit problems where passive arms incur costs rather than active arms generating rewards are often referred to as the tax problems.

The Gittins index can be extended, quite straightforwardly, to an arm with a non-zero reward function $\overset{\circ}{r}(s)$ in passivity. Recall that when $\overset{\circ}{r}(s) = 0$ for all $s \in \mathcal{S}$, the Gittins index is the maximum reward per discounted time attainable by activating the arm at least once ($\tau \geq$

1) and then stopping at an optimal time τ^* (see (2.20)). With a non-zero $\overset{\circ}{r}(s)$, the index, measuring the value for activating the arm at a particular state, should be given by the maximum *additional* reward per discounted time attainable by activating the arm up to an optimal stopping time $\tau^* \geq 1$ as compared to being passive all through. We thus have

$$\nu(s) = \max_{\tau\geq 1} \frac{\mathbb{E}\left[\sum_{t=1}^{\tau}\beta^{t-1}r(S(t)) \,\middle|\, S(1)=s\right] + \frac{1}{1-\beta}\mathbb{E}\left[\beta^{\tau}\overset{\circ}{r}(S(\tau)) - \overset{\circ}{r}(s) \,\middle|\, S(1)=s\right]}{\mathbb{E}\left[\sum_{t=1}^{\tau}\beta^{t-1} \,\middle|\, S(1)=s\right]}. \tag{3.26}$$

A more intuitive interpretation of this index form is the subsidy-for-passivity concept: $\nu(s)$ is the minimum subsidy for passivity needed to make the active and passive actions equally attractive at state s. Specifically, under a subsidy for passivity given by the index $\nu(s)$, the total discounted reward for activating the arm until a stopping time τ is upper bounded by that of being passive all through:

$$\begin{aligned}\sum_{t=1}^{\infty}\beta^{t-1}(\nu(s)+\overset{\circ}{r}(s)) \quad \geq \quad & \mathbb{E}\left[\sum_{t=1}^{\tau}\beta^{t-1}r(S(t)) \,\middle|\, S(1)=s\right] \\ & +\mathbb{E}\left[\sum_{t=\tau+1}^{\infty}\beta^{t-1}(\overset{\circ}{r}(S(\tau))+\nu(s)) \,\middle|\, S(1)=s\right],\end{aligned} \tag{3.27}$$

where equality holds for the optimal stopping time τ^*. We thus arrive at (3.26).

3.4.2 BANDITS WITH SWITCHING COST AND SWITCHING DELAY

Assumption A3.3 is necessary for the index theory. It is not difficult to show that bandits with switching cost are special cases of the restless bandit model.

Suppose that a setup cost $c_i(s)$ is incurred when switching to arm i that is in state s. A more general formulation includes, in addition to the setup cost, a tear-down cost dependent of the state of the arm being switching away from. Banks and Sundaram, 1994 [29] showed that the tear-down cost can be absorbed into the setup cost based on the simple fact that a setup cost $c_i(s)$ of arm i in state s must follow a tear-down cost of the same arm in the same state at some earlier time instant. It is thus without loss of generality to assume only setup costs.

The bandit problem with switching cost can be formulated as a restless bandit (without switching cost) by augmenting the state of arm i at time t with the active/passive action $a_i(t-1) \in \{0,1\}$ at time $t-1$. The reward $r_i'(s(t), a_i(t-1))$ for activating arm i is thus given by

$$r_i'(s, a_i(t-1)=1) = r_i(s), \quad r_i'(s, a_i(t-1)=0) = r_i(s) - c_i(s), \tag{3.28}$$

where $r_i(s)$ is the reward function for arm i in the bandit problem with switching cost. Passive arms generate no reward. The augmented state, however, becomes restless since arm i in state $(s, a_i(t-1)=1)$ at time t, when made passive, transits deterministically to state $(s, a_i(t)=0)$ at time $t+1$.

This is a special restless bandit problem due to the specific and deterministic state transition in passivity. It has been shown by Glazebrook and Ruiz-Hernandez, 2005 [95] that this restless bandit problem is indexable and the Whittle index can be efficiently computed via an adaptive greedy algorithm. The optimality of the Whittle index, however, does not hold in general. Sundaram, 2005 [186] has shown that no strongly decomposable index policies offer optimality for a general bandit with switching cost. A number of studies exist that either provide sufficient conditions for index-type optimal policies or consider special cases of the problem. See a comprehensive survey by Jun, 2004 [104].

Another form of switching penalty is in terms of a startup delay when re-initiating the work on a project. The startup delay is modeled as a random variable dependent on the state of the project being re-initiated. While imposing no direct switching cost, a delay affects the future return through the discounting: all future rewards are discounted with an extra factor determined by the state-dependent random delay. Similar to the case with switching cost, bandits with switching delays can be formulated under the restless bandit model. The resulting bandit model, however, becomes semi-Markovian. Studies on bandits with switching delays and switching costs can be found in Asawa and Teneketzis, 1996 [18] and Nino-Mora, 2008 [159].

3.5 VARIATIONS IN PERFORMANCE MEASURE

3.5.1 STOCHASTIC SHORTEST PATH BANDIT

It is a well-established fact that maximizing the discounted reward with a discount factor β over an infinite horizon is equivalent to maximizing the total *undiscounted* reward over a random horizon with a geometrically distributed length T given by $\mathbb{P}[T = n] = (1 - \beta)\beta^{n-1}$ for $n = 1, 2, \ldots$ (see, for example, Puterman, 2005 [164]). The value of T is independent of the policy deployed. The index theory thus directly apply to this type of random horizon problems.

In this section, we show that the index theory extends to more complex random horizon problems in which the horizon length depends on the policy. Quite often, the objective is to maximize or minimize the expected horizon length. Such random horizon problems arise in applications concerned with finding the most rewarding (or the least costly) path to a terminating goal such as finding an object or accomplishing a task. The resulting bandit model is referred to as the *stochastic shortest path bandit*, where the word "shortest" conforms to the convention that edge/vertex weights represent costs and the word "stochastic" captures the random walk nature of the problem in which our action only specifies *probabilistically* the next edge to traverse.

The stochastic shortest path bandit model: Before presenting the formal definition, we consider the following illuminating example from Dumitriu, Tetali, and Winkler, 2003 [75].

Example 3.17 Consider a line graph with vertices $0, 1, \ldots, 5$ as shown in Figure 3.4. There are two tokens initially at vertices 2 and 5, respectively. A valuable gift at vertex 3 can be collected if either token reaches 3. At any time, you may pay $1 for a random move made by a token of your

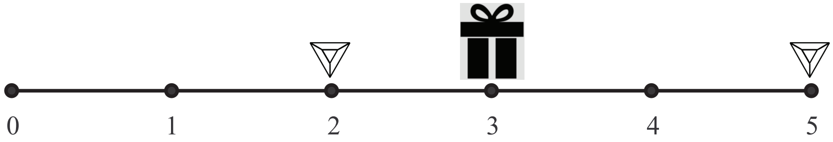


Figure 3.4: A stochastic shortest path bandit problem.

choice. The chosen token moves to one of its neighbors with equal probability (this probability is 1 if it has a single neighbor). What is the optimal strategy for moving the tokens?

For this simple problem, it is not difficult to see that the optimal strategy is to move the token at 2 first and then switch permanently to the other token if the game does not end immediately. The resulting expected total cost can be calculated as follows. Let x denote the expected total cost for reaching the prize by moving the token located at vertex 5. It satisfies the following equation:

$$x = 1 + \frac{1}{2} \times 1 + \frac{1}{2} \times (1 + x), \tag{3.29}$$

which leads to $x = 4$. The total expected cost of the strategy given above is thus $\frac{1}{2} \times 1 + \frac{1}{2}x = 3$.

We now formally define the stochastic shortest path bandit model. For consistency and without loss of generality, we pose the problem as maximizing reward rather than minimizing cost.

Definition 3.18 *The Stochastic Shortest Path Bandit Model:*
Consider N arms, each with state space $\mathcal{S}_i$ $(i = 1, \ldots, N)$. At time $t = 1, 2, \ldots$, based on the observed states $[s_1(t), \ldots, s_N(t)]$ of all arms, one arm, say arm i, is activated. The active arm offers a reward $r_i(s_i(t))$ dependent of its current state and changes state according to a Markov transition rule $p_i(s, s')$ $(s, s' \in \mathcal{S}_i)$. The states of all passive arms remain frozen. Each arm has a terminating state Δ satisfying $r_i(\Delta) = 0$ and $p_i(\Delta, \Delta) = 1$. The objective is to maximize the expected total (undiscounted) reward until one arm reaches its terminating state for the first time. We assume here that the expected number of steps to reach the terminating state from every state $s \in \mathcal{S}_i$ is finite for all arms.

This bandit model appeared in the literature in various special forms. The general formulation was considered by Dumitriu, Tetali, and Winkler, 2003 [75], under a vividly descriptive title of "Playing Golf with Two Balls." The formulation was given in terms of minimizing cost.

The extended Gittins index: We take the restart-in-state characterization of the Gittins index and generalize it to the stochastic shortest path bandit model.

Consider a single arm and omit the subscript for the arm label. The index $\nu(s)$ for state s is the maximum expected total reward accrued starting in state s and with the option of restarting in state s at every decision time until the arm reaches its terminating state. In other words, whenever we end up in a state less advantageous (in terms of the total reward until termination) than the initial state s, we go back to s and restart. The decision problem is to find the optimal set of such states for restarting. Let $\mathcal{Q} \subset \mathcal{S}$ denote a subset of states in which we restart to s. The resulting Markov reward process essentially aggregates all states in $\mathcal{Q}$ with state s. Its state space is given by $\tilde{\mathcal{S}} = \mathcal{S} \setminus \mathcal{Q}$, and the transition probabilities $\tilde{p}(u, u')$ are

$$\tilde{p}(u, u') = \begin{cases} p(u, u') & \text{if } u, u' \in \tilde{\mathcal{S}}, u' \neq s \\ \sum_{w \in \mathcal{Q} \cup \{s\}} p(u, w) & \text{if } u \in \tilde{\mathcal{S}}, u' = s \end{cases} . \tag{3.30}$$

The cumulative reward until absorption in Δ of this Markov chain gives us the expected total reward $V_{\mathcal{Q}}(s)$ until termination under the restart set $\mathcal{Q}$. The index $\nu(s)$ is then obtained by optimizing the restart set $\mathcal{Q}$:

$$\nu(s) = \max_{\mathcal{Q} \subset \mathcal{S}} V_{\mathcal{Q}}(s). \tag{3.31}$$

Let $\mathcal{Q}^*(s)$ denote the optimal restart set that attains the index value $\nu(s)$. Similar to Property 2.5, we have

$$\left\{s' : \nu(s') < \nu(s)\right\} \subseteq \mathcal{Q}^*(s) \subseteq \left\{s' : \nu(s') \leq \nu(s)\right\}, \tag{3.32}$$

which states the intuitive: the optimal restart set (states less advantageous than the initial state s) for initial state s consists of states with indices smaller than that of s. This also leads to a recursive algorithm, similar to that described in Section 2.5, for computing the indices one state at a time, starting with the state with the greatest index value.

The following example, first posed by Bellman, 1957 [31], and later used by Kelly, 1979 [113], to illustrate the index theorem, fits perfectly under the stochastic shortest path bandit model.

Example 3.19 *The Gold-Mining Problem:*
There are N gold mines, each containing initially x_i $(i = 1, 2, \ldots, N)$ tons of gold. There is a single gold-mining machine. Each time the machine is used at mine i, there is a probability q_i that it will mine a fraction η_i of the remaining gold at the mine and remain in working order and a probability $1 - q_i$ that it will mine no gold and be damaged beyond repair. The objective is to maximize the expected amount of gold mined before the machine breaks down.

The stochastic shortest path bandit formulation of the problem is quite straightforward: each mine corresponds to an arm with the state being the amount of the remaining gold and a terminating state Δ denoting the machine breaking down. When arm i is activated in state s_i (i.e., the machine is used at mine i), the reward and the transition probabilities are given by

$$r_i(s_i) = q_i \eta_i s_i \tag{3.33}$$

$$p_i(s_i, s_i') = \begin{cases} q_i & \text{if } s_i' = (1 - \eta_i)s_i \\ 1 - q_i & \text{if } s_i' = \Delta \\ 0 & \text{otherwise} \end{cases} . \tag{3.34}$$

Since the state and the associated reward decrease with each mining, it is easy to see that in the restart-in-state problem, the optimal action is to restart in state s_i right after each mining if the machine has not broken down. The index $\nu_i(s_i)$, the expected total reward obtained with restart until the machine breaks down, is given by

$$\nu_i(s_i) = r_i(s_i) + q_i \nu_i(s_i), \tag{3.35}$$

which leads to

$$\nu_i(s_i) = \frac{q_i \eta_i s_i}{1 - q_i}. \tag{3.36}$$

The optimal policy is to assign the machine to the mine $i^* = \arg\max_{i=1,\dots,N} \nu_i(s_i)$ with the greatest current index value.

3.5.2 AVERAGE-REWARD AND SENSITIVE-DISCOUNT CRITERIA

The average-reward criterion can be rather unselective. Assume that the states of each arm communicate well enough that the average reward does not depend on the initial state. If the objective is to simply maximize the average reward, also referred to as the gain denoted by g, a simple policy that activates, at each t, the arm with the highest limiting reward rate suffices. Specifically, let $\{\gamma_i(s)\}_{s \in \mathcal{S}_i}$ denote the unique stationary distribution of arm i. The gain g_i (i.e., the limiting reward rate) of arm i is given by

$$g_i = \sum_{s \in \mathcal{S}_i} r(s) \gamma_i(s). \tag{3.37}$$

The policy that plays a fixed arm $i^* = \arg\max_{i=1,\dots,N} g_i$ all through suffices to maximize the average reward.

A stronger optimality criterion is to require both gain optimal and bias $h(\mathbf{s})$ optimal, where the bias function $h(\mathbf{s})$ is the differential reward caused by the transient effect of starting at state $\mathbf{s} = [s_1, \dots, s_N]$ rather than the stationary distributions of the arms. The following modified Gittins index leads to an optimal policy under this criterion:

$$\nu_i(s) = \max_{\tau \geq 1} \frac{\mathbb{E}\left[\sum_{t=1}^{\tau} r_i(S_i(t)) \;\middle|\; S_i(1) = s\right]}{\mathbb{E}\left[\tau \mid S_i(1) = s\right]}. \tag{3.38}$$

Other stronger optimality criteria such as Veinott optimality (Veinott, 1966 [202]) and Blackwell optimality (Blackwell, 1962 [41]) (i.e., simultaneous maximization of total-discounted-reward under all discount factors sufficiently close to 1) were considered in the bandit setting by Katehakis and Rothblum, 1996 [111]. It was shown that the decomposition structure of the optimal policy is preserved under these criteria and can be computed through a Laurent expansion of the Gittins index (as the value under restarting) around $\beta = 1$.

3.5.3 FINITE-HORIZON CRITERION: BANDITS WITH DEADLINES

A finite-horizon bandit model under the total-reward criterion (see (2.2)) can be generalized to include cases where projects/arms have different deadlines for completion. We have discussed in Section 3.1.4 that the index theory does not hold for a finite-horizon bandit, since the maximum reward rate promised by the Gittins index may only be realized by a stopping time that exceeds the remaining time horizon. This leads to a natural question on whether a modified Gittins index resulted from searching over a constrained set of stopping times determined by the remaining time horizon might offer an optimal solution. The answer, unfortunately, is negative as discussed by Gittins and Jones, 1974 [87].

The finite-horizon problem can, however, be formulated as a restless bandit problem by augmenting the state of each project with the remaining time to the deadline. This leads to a special class of restless bandit problems where the state of a passive arm changes in a deterministic way (i.e., the remaining time to the deadline is reduced by one at each time). The indexability and priority index policies for this class of restless bandit problems were studied by Nino-Mora, 2011 [160].

CHAPTER 4

Frequentist Bandit Model

In this chapter, we introduce basic formulations and major results of the bandit problems within the frequentist framework. Toward the end of this section, we draw connections between the Bayesian and the frequentist bandit models from an algorithmic perspective. In particular, we discuss how learning algorithms developed within one framework can be applied and evaluated in another. A success story is the Thompson Sampling algorithm. Originally developed from a Bayesian perspective (see Example 2.3), it has recently gain wide popularity as a solution to frequentist bandit problems and has enjoyed much fame due to its strong performance under frequentist criteria as demonstrated both empirically and analytically.

4.1 BASIC FORMULATIONS AND REGRET MEASURES

Definition 4.1 *The Frequentist Multi-Armed Bandit Model:*
Consider an N-armed bandit and a single player. At each time $t = 1, \ldots, T$, the player chooses one arm to play. Successive plays of arm i $(i = 1, \ldots, N)$ generate independent and identically distributed random rewards $X_i(t)$ drawn from an unknown distribution $F_i(x)$ with a finite (unknown) mean μ_i. The objective is to maximize the expected value of the total reward accrued over T plays.

An arm-selection policy π is a sequence of functions $\{\pi_1, \ldots, \pi_T\}$, where π_t maps from the player's past actions and reward observations to the arm to be selected at time t. Let $V_\pi(T; \mathbf{F})$ denote the total expected reward obtained over T plays under policy π for a given configuration of the reward distributions $\mathbf{F} = (F_1, \ldots, F_N)$. With a slight abuse of notation, let π_t denote the arm chosen at time t under policy π, and $X_{\pi_t}(t)$ the random reward obtained at time t. We have

$$\begin{aligned} V_\pi(T; \mathbf{F}) &= \mathbb{E}_\pi\left[\sum_{t=1}^{T} X_{\pi_t}(t)\right] && (4.1) \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{E}_\pi\left[X_{\pi_t}(t)\mathbb{I}(\pi_t = i)\right] && (4.2) \\ &= \sum_{i=1}^{N} \mu_i \mathbb{E}_\pi\left[\tau_i(T)\right], && (4.3) \end{aligned}$$

where $\tau_i(T) = \sum_{t=1}^{T} \mathbb{I}(\pi_t = i)$ is the total number of times that policy π chooses arm i, whose dependence on $\mathbf{F}$ is omitted in the notation for simplicity. Note that (4.3) follows from Wald's identity.

4.1.1 UNIFORM DOMINANCE VS. MINIMAX

It is clear from (4.3) that the expected total reward $V_\pi(T; \mathbf{F})$ obtained under a policy π depends on the unknown arm configuration $\mathbf{F}$. This brings two issues if $V_\pi(T; \mathbf{F})$ is adopted as a measure for the performance of a policy π.

The first issue is how to compare policies and define optimality. A policy that always plays arm 1 would work perfectly, achieving the maximum value of $V_\pi(T; \mathbf{F})$ for arm configurations satisfying $\mu_1 = \max_i\{\mu_i\}$. This also makes characterizing the fundamental performance limit a trivial task: $\max_i\{\mu_i\}\, T$ is the achievable upper bound on $V_\pi(T; \mathbf{F})$ under every given $\mathbf{F}$.

It goes without saying that such heavily biased strategies that completely forgo learning are of little interest. They should be either excluded from consideration or penalized when measuring their performance. The former leads to the uniform-dominance approach to comparing policies and defining optimality, the latter the minimax approach.

In the uniform-dominance approach, the objective is a policy whose performance uniformly (over all possible arm configurations $\mathbf{F}$ in a certain family $\mathcal{F}$ of distributions) dominates that of others. Such a policy, while naturally fits the expectation for optimal policies, does not exist in general without a proper restriction on the class of policies being considered. More specifically, the bandit formulation under the uniform-dominance approach involves specifying two sets: the class Π of admissible policies (which an optimal policy needs to dominate) and the set $\mathcal{F}^N$ of all possible arm configurations (over which the dominance needs to hold uniformly). The more general these two sets are, the stronger the optimality statement.

In the minimax approach, every policy is admissible. The performance of a policy, however, is measured against the worst possible arm configuration $\mathbf{F}$ specific to this policy and the horizon length T. The worst-case performance of a policy no longer depends on $\mathbf{F}$. Meaningful comparison of all strategies can be carried out and optimality can be rigorously defined. This approach can also be viewed through the lens of a two-player zero-sum game where the arm configuration $\mathbf{F}$ is chosen by nature who plays the role of an opponent.

This brings us to the second issue with using the total expected reward $V_\pi(T; \mathbf{F})$ as the performance measure. The total expected reward is limited by the arm configuration $\mathbf{F}$. A policy effective at learning the arm characteristics may still return low utility in terms of the total reward accrued simply due to the given arm configuration. In particular, using reward as the payoff in the minimax approach leads to a trivial game: the action of nature is to set $\{\mu_i\}$ as small as possible, and the action of the player is to choose the arm for which the smallest possible mean is the greatest.

A modified objective is thus necessary, one that measures the efficacy of a policy in *learning* the unknown arm configuration $\mathbf{F}$ rather than the absolute return of rewards. A natural candidate is the *difference* in the total expected reward against an oracle who knows the arm configuration $\mathbf{F}$ and plays optimally using that knowledge. This leads to the notion of *regret* as discussed below.

4.1.2 PROBLEM-SPECIFIC REGRET AND WORST-CASE REGRET

Suppose that $\mathbf{F}$ is known. The optimal action is obvious: choose, at each time, the arm with the greatest mean value:

$$\mu_* = \max\{\mu_1, \ldots, \mu_N\}. \tag{4.4}$$

The omniscient oracle enjoys the maximum expected reward μ_* per play and the maximum expected total reward $\mu_* T$ over the horizon of length T. When $\mathbf{F}$ is unknown, it is inevitable that an inferior arm with a mean less than μ_* is played from time to time. The question is then how much the loss is as compared to the maximum expected total reward $\mu_* T$ in the case of known $\mathbf{F}$. This cumulative reward loss is referred to as *regret* or *cost of learning* and is given by

$$\begin{aligned} R_\pi(T;\mathbf{F}) &= \mu_* T - V_\pi(T;\mathbf{F}) && (4.5)\\ &= \sum_{i:\mu_i<\mu_*} (\mu_* - \mu_i)\mathbb{E}_\pi\left[\tau_i(T)\right], && (4.6)\end{aligned}$$

where (4.6) follows from (4.3) and that $\sum_{i=1}^N \mathbb{E}_\pi[\tau_i(T)] = T$.

In the uniform-dominance approach, a policy π in the class Π of admissible policies is optimal if, for all admissible policies $\pi' \in \Pi$ and all arm configurations $\mathbf{F} \in \mathcal{F}^N$, we have

$$R_\pi(T;\mathbf{F}) \leq R_{\pi'}(T;\mathbf{F}). \tag{4.7}$$

We refer to $R_\pi(T;\mathbf{F})$ as the *problem-specific regret* to highlight its dependence on $\mathbf{F}$ and differentiate it from the worst-case regret defined below.

In the minimax approach, the *worst-case regret* of a policy π is defined as

$$\bar{R}_\pi(T;\mathcal{F}) \triangleq \sup_{\mathbf{F}\in\mathcal{F}^N} R_\pi(T;\mathbf{F}). \tag{4.8}$$

Note that the worst-case regret $\bar{R}_\pi(T;\mathcal{F})$ is no longer a function of $\mathbf{F}$, but a function of the set $\mathcal{F}$ of possible reward distributions. We often omit this dependence in the notation when there is no ambiguity. A policy π is minimax optimal if, for all π', we have

$$\bar{R}_\pi(T;\mathcal{F}) \leq \bar{R}_{\pi'}(T;\mathcal{F}).$$

A policy can be evaluated under both the uniform-dominance approach and the minimax approach. It holds trivially that its problem-specific regret is no greater than its worst-case regret. Specifically, for all $\mathbf{F} \in \mathcal{F}^N$, we have

$$R_\pi(T;\mathbf{F}) \leq \bar{R}_\pi(T;\mathcal{F}).$$

As we will see in subsequent sections, for the same policy, these two regret measures may have drastically different scaling behaviors in T. It is also natural to expect that optimality under one regret measure need not transfer to the other, although policies that are order optimal under both measures do exist.

Of particular interest is the rate at which $R_\pi(T;\mathbf{F})$ and $\bar{R}_\pi(T;\mathcal{F})$ grow with T. A sublinear regret growth rate implies that the maximum expected reward μ_* per play can be approached as the learning horizon T tends to infinite. Thus, all policies offering a sublinear regret order are optimal in terms of maximizing the long-term average reward. Regret, however, is a finer performance measure than the long-term time average. It not only indicates whether the long-term average reward converges to μ^*, but also measures the rate of the convergence. For example, a policy with $R_\pi(T;\mathbf{F}) \sim O(\log T)$, while converging to the same expected reward per play as a policy with $R_{\pi'}(T;\mathbf{F}) \sim O(\sqrt{T})$, is far more superior in terms of learning efficiency.

A policy π is *asymptotically optimal* in terms of its problem-specific regret if for all $\mathbf{F} \in \mathcal{F}^N$,

$$\liminf_{T\to\infty} \frac{R_\pi(T;\mathbf{F})}{\inf_{\pi'\in\Pi} R_{\pi'}(T;\mathbf{F})} = 1. \tag{4.9}$$

A policy π is *order optimal* if for all $\mathbf{F} \in \mathcal{F}^N$,

$$\liminf_{T\to\infty} \frac{R_\pi(T;\mathbf{F})}{\inf_{\pi'\in\Pi} R_{\pi'}(T;\mathbf{F})} < C(\mathbf{F}) \tag{4.10}$$

for some constant $C(\mathbf{F})$ independent of T, but dependent of $\mathbf{F}$ in general.

Asymptotic optimality under the minimax regret criterion can be similarly defined. For order optimality, however, the constant C corresponding to that in (4.10) is no longer a function of the arm configuration $\mathbf{F}$. Specifically, a policy π is minimax order optimal if there exists a constant C such that

$$\liminf_{T\to\infty} \frac{\bar{R}_\pi(T;\mathcal{F})}{\inf_{\pi'} \bar{R}_{\pi'}(T;\mathcal{F})} < C. \tag{4.11}$$

4.1.3 REWARD DISTRIBUTION FAMILIES AND ADMISSIBLE POLICY CLASSES

Under both the uniform-dominance approach and the minimax approach, the family of all possible reward distributions needs to be specified. For the uniform dominance approach, the class of admissible policies also need to be specified. Below we discuss commonly adopted distribution families and policy classes.

Family of reward distributions: Often, the specific application under consideration imposes natural restrictions on the reward distributions that the arms may assume. For example, in the clinical trial application with dichotomous responses, the random rewards from each arm are known to follow a Bernoulli distribution. There is no reason to require dominance under distributions outside this family of distributions, since generality in models often comes with a sacrifice in targeted performance.

The family $\mathcal{F}$ of possible reward distributions depends on the specific application and available prior knowledge. General approaches to specifying $\mathcal{F}$ fall under two categories: parametric and nonparametric. In the parametric setting, $\mathcal{F}$ is defined by a known distribution type $F(x;\theta)$ with an unknown parameter θ belonging to a known set Θ:

$$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}. \tag{4.12}$$

Under this setting, all arms assume the same type of distribution and differ only in the parameter θ. An example is the clinical trial application where all arms assume a Bernoulli distribution with an unknown mean $\theta \in [0, 1]$. The parameter θ can be a vector to allow multivariate distributions, for example, a Gaussian distribution parameterized by mean and variance.

A nonparametric characterization of $\mathcal{F}$ does not stipulate a specific distribution, but rather imposing certain conditions, often on the concentration behavior of the distributions. Commonly considered in the literature are—from the most concentrated to the least concentrated—the set of distributions with bounded support, sub-Gaussian distributions, light-tailed distributions, and heavy-tailed distributions. Specifying $\mathcal{F}$ by imposing constraints on the concentration behavior is quite natural. A less concentrated distribution leads to less accurate inference of its mean from random samples, thus more resistance in terms of achievable regret performance and more demanding in terms of learning algorithm design.

Class of admissible policies: A natural way to define the class Π of admissible policies is to consider strategies that offer good performance for all arm configurations in $\mathcal{F}^N$, hence excluding heavily biased policies.

Definition 4.2 A policy π is *consistent* if for all $\mathbf{F} \in \mathcal{F}^N$, it asymptotically achieves the highest average reward per play μ^*:

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[V_\pi(T;\mathbf{F})\right]}{T} = \mu^*. \tag{4.13}$$

Equivalently, a consistent policy π experiences diminishing regret per play as T tends to infinity:

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[R_\pi(T;\mathbf{F})\right]}{T} = 0. \tag{4.14}$$

This is often referred to as Hannan consistency originally introduced in an adversarial game setting (Hannan, 1957 [98]).

For a given $\alpha \in (0, 1)$, a policy is α-*consistent* if for all $\mathbf{F} \in \mathcal{F}^N$, it satisfies

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[R_\pi(T;\mathbf{F})\right]}{T^\alpha} = 0. \tag{4.15}$$

A policy that is α-consistent for all $\alpha \in (0, 1)$ is called *uniformly good*.

Let Π_c, Π_α, and Π_u denote, respectively, the classes of consistent, α-consistent, and uniformly good policies. It is easy to see that $\Pi_u \subset \Pi_\alpha \subset \Pi_c$ for all $\alpha \in (0, 1)$.

4.2 LOWER BOUNDS ON REGRET

In this section, we establish lower bounds on achievable regret under both the uniform-dominance and minimax formulations. These lower bounds serve as benchmarks for determining the asymptotic or order optimality of learning policies.

4.2.1 THE PROBLEM-SPECIFIC REGRET

Recall that $X_i(t)$ denotes the random reward offered by arm i at time t. In the parametric setting, $X_i(t)$ is drawn from a univariate cumulative distribution function $F(x;\theta_i)$, where $F(\cdot;\cdot)$ is known and the parameter θ_i is unknown belonging to a parameter space Θ. For notation simplicity, we assume that $F(x;\theta_i)$ is differentiable and work with the corresponding density function $f(x;\theta_i)$.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ denote the vector of the unknown parameters. Let $\mu(\theta_i) = \int_{x=-\infty}^{\infty} x f(x;\theta_i) \mathrm{d}x$ denote the expected reward of arm i. Let $\mu_* = \max_{i=1,\ldots,N} \mu(\theta_i)$ and $\theta_* = \arg\max_{i=1,\ldots,N} \mu(\theta_i)$. The regret $R_\pi(T;\boldsymbol{\theta})$ of policy π can be written as

$$R_\pi(T;\boldsymbol{\theta}) = \sum_{i:\mu(\theta_i)<\mu_*} (\mu_* - \mu(\theta_i))\, \mathbb{E}_\pi[\tau_i(T)], \tag{4.16}$$

where we have replaced the notation $R_\pi(T;\mathbf{F})$ by $R_\pi(T;\boldsymbol{\theta})$ to make explicit the dependency on $\boldsymbol{\theta}$ in the parametric setting.

Let $D(\theta||\lambda) := D(f(x;\theta)||f(x;\lambda))$ denote the Kullback–Leibler (KL) divergence between $f(x;\theta)$ and $f(x;\lambda)$ defined as

$$D(\theta||\lambda) = \mathbb{E}_{X\sim f(x;\theta)}\left[\log \frac{f(X;\theta)}{f(X;\lambda)}\right] = \int_{x=-\infty}^{\infty} f(x;\theta) \log \frac{f(x;\theta)}{f(x;\lambda)} \mathrm{d}x. \tag{4.17}$$

Specifically, the KL divergence $D(\theta||\lambda)$ is the expected value of the log-likelihood ratio of a random sample X with respect to the two distributions $f(x;\theta)$ and $f(x;\lambda)$ where the random sample X is drawn from the first distribution $f(x;\theta)$. It measures the rate (in terms of the number of samples) at which the two distributions can be distinguished based on samples drawn from the first distribution. The smaller the KL divergence, the harder it is to distinguish the two distributions. KL divergence is nonnegative (with zero value attained if and only if the two distributions are identical) and generally asymmetric in the two distributions.

We assume the following regularity conditions on the univariate distributions and the parameter space.

Assumption 4.3

A4.1 (*Denseness of* Θ:) For all $\delta > 0$ and $\theta \in \Theta$, there exists $\theta' \in \Theta$ such that $\mu(\theta) < \mu(\theta') < \mu(\theta) + \delta$.

A4.2 (*Identifiability via the mean:*) The univariate distribution function $F(\cdot;\cdot)$ is such that for all $\theta, \lambda \in \Theta$, if $\mu(\theta) < \mu(\lambda)$, then $0 < D(\theta||\lambda) < \infty$.

A4.3 (*Continuity of* $D(\theta||\lambda)$ *in* λ:) For all $\epsilon > 0$ and $\theta, \lambda \in \Theta$ with $\mu(\theta) < \mu(\lambda)$, there exists $\delta > 0$ such that $|D(\theta||\lambda) - D(\theta||\lambda')| < \epsilon$ whenever $\mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta$.

Theorem 4.4 Lower Bound on Problem-Specific Regret in Univariate Parametric Setting: *Under the regularity conditions A4.1–A4.3, for every α-consistent policy π and every arm configuration $\boldsymbol{\theta}$ such that $\mu(\theta_i)$ are not all equal, we have*

$$\liminf_{T\to\infty} \frac{R_\pi(T;\boldsymbol{\theta})}{\log T} \geq (1-\alpha) \sum_{i:\mu(\theta_i)<\mu_*} \frac{\mu_* - \mu(\theta_i)}{D(\theta_i||\theta_*)}. \tag{4.18}$$

For every uniformly good policy, we have

$$\liminf_{T\to\infty} \frac{R_\pi(T;\boldsymbol{\theta})}{\log T} \geq \sum_{i:\mu(\theta_i)<\mu_*} \frac{\mu_* - \mu(\theta_i)}{D(\theta_i||\theta_*)}. \tag{4.19}$$

Theorem 4.4 implies that to achieve uniformly good performance over all arm configurations, each suboptimal arm with $\mu(\theta_i) < \mu_*$ needs to be explored, asymptotically, no fewer than $\frac{1}{D(\theta_i||\theta_*)} \log T$ times, which grows in a logarithmic order with T and is inversely proportional to the distribution divergence $D(\theta_i||\theta_*)$ between this suboptimal arm and the optimal arm.

Proof. The proof is based on the following key lemma that establishes a high-probability lower bound on the number of times that every suboptimal arm has to be explored to ensure α-consistency.

Lemma 4.5 *Consider an arbitrary α-consistent policy and anbitrary $\boldsymbol{\theta} \in \Theta^N$. For every suboptimal arm i with $\mu(\theta_i) < \mu_*$, we have, for all $\epsilon > 0$,*

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\theta}} \left[\tau_i(T) \geq (1-\epsilon) \frac{(1-\alpha)\log T}{D(\theta_i||\theta_*)} \right] = 1, \tag{4.20}$$

where $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the probability measure induced by $\boldsymbol{\theta}$ with the dependence on the policy omitted from the notation.

To prove the lemma, assume that, without loss of generality, $\mu_* = \mu(\theta_1)$ and $\mu(\theta_2) < \mu_*$ (i.e., arm 1 is an optimal arm, not necessarily unique, and arm 2 a suboptimal arm). It suffices to prove (4.20) for $i = 2$.

Fix $0 < \delta < \epsilon$. Based on A4.1–A4.3, there exists $\theta_2' \in \Theta$ such that $\mu(\theta_2') > \mu_*$ and

$$0 < D(\theta_2||\theta_2') - D(\theta_2||\theta_1) < \delta D(\theta_2||\theta_1). \tag{4.21}$$

It thus suffices to show

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}\right] = 0, \tag{4.22}$$

where $c_1 = (1-\epsilon)(1+\delta)$ satisfying $c_1 < 1$. In other words, to distinguish arm configuration $\boldsymbol{\theta}$ under which arm 1 is optimal from that of $\boldsymbol{\theta}' = [\theta_1, \theta_2', \theta_3, \ldots, \theta_N]$ under which arm 2 is optimal, the probability that an α-consistent policy plays arm 2 less than $c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}$ times has to diminish to zero.

Let $X(1), X(2), \ldots$ be successive observations from arm 2, where for notation simplicity, we have omitted the arm index and labeled the reward samples from arm 2 with consecutive integers. Define the sum log-likelihood ratio of m samples from arm 2 with respect to the two distributions parameterized by θ_2 and θ_2', respectively:

$$L_m = \sum_{j=1}^{m} \log \frac{f(X(j);\theta_2)}{f(X(j);\theta_2')}. \tag{4.23}$$

Choose $c_2 \in (c_1, 1)$. We have

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}\right] &= \mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} > c_2(1-\alpha)\log T\right] \\ &\quad + \mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} \le c_2(1-\alpha)\log T\right].\end{aligned} \tag{4.24}$$

Next, we show each of the two terms on the right-hand side of the above equation tends to zero as T approaches infinity. For the first term, we have

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} > c_2(1-\alpha)\log T\right] \\ &\le \mathbb{P}_{\boldsymbol{\theta}}\left[\max_{m \le c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}} L_m > c_2(1-\alpha)\log T\right] &(4.25) \\ &= \mathbb{P}_{\boldsymbol{\theta}}\left[\max_{m \le c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}} \frac{L_m}{c_1(1-\alpha)\log T / D(\theta_2||\theta_2')} > \frac{c_2}{c_1} D(\theta_2||\theta_2')\right]. &(4.26)\end{aligned}$$

By the strong law of large numbers, $\frac{1}{m}L_m$ converges to $D(\theta_2||\theta_2')$ almost surely under $\boldsymbol{\theta}$, which implies that $\frac{1}{m}\max_{k\le m} L_k$ converges to $D(\theta_2||\theta_2')$ almost surely. Noticing that c_2/c_1 is bounded away from 1, we conclude, from (4.26), that

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1 \frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} > c_2(1-\alpha)\log T\right] = 0. \tag{4.27}$$

We now consider the second term on the right-hand side of (4.24):

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} \leq c_2(1-\alpha)\log T\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{I}\left(\tau_2(T) < c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} \leq c_2(1-\alpha)\log T\right)\right] && (4.28)\\
&= \mathbb{E}_{\boldsymbol{\theta}'}\left[\mathbb{I}\left(\tau_2(T) < c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} \leq c_2(1-\alpha)\log T\right)e^{L_{\tau_2}}\right] \quad \text{(Change of measure)} && (4.29)\\
&\leq \mathbb{E}_{\boldsymbol{\theta}'}\left[\mathbb{I}\left(\tau_2(T) < c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} \leq c_2(1-\alpha)\log T\right)e^{c_2(1-\alpha)\log T}\right] && (4.30)\\
&= T^{c_2(1-\alpha)}\mathbb{P}_{\boldsymbol{\theta}'}\left[\tau_2(T) < c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} \leq c_2(1-\alpha)\log T\right] && (4.31)\\
&\leq T^{c_2(1-\alpha)}\mathbb{P}_{\boldsymbol{\theta}'}\left[\tau_2(T) < c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}\right] && (4.32)\\
&= T^{c_2(1-\alpha)}\mathbb{P}_{\boldsymbol{\theta}'}\left[T-\tau_2(T) \geq T - c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}\right] && (4.33)\\
&\leq T^{c_2(1-\alpha)}\frac{\mathbb{E}_{\boldsymbol{\theta}'}[T-\tau_2(T)]}{T-c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}} \quad \text{(Markov inequality)} && (4.34)\\
&\leq \frac{T^{c_2(1-\alpha)}T^{\alpha}}{T-c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')}} \quad (\alpha\text{-consistency}), && (4.35)
\end{aligned}$$

where (4.35) follows from the fact that under arm configuration $\boldsymbol{\theta}'$ where arm 2 is the unique optimal arm, an α-consistent policy plays the $N-1$ suboptimal arms fewer than T^{α} times in expectation.

Since $c_2 < 1$, the above leads to

$$\lim_{T\to\infty}\mathbb{P}_{\boldsymbol{\theta}}\left[\tau_2(T) < c_1\frac{(1-\alpha)\log T}{D(\theta_2||\theta_2')},\ L_{\tau_2} \leq c_2(1-\alpha)\log T\right] = 0. \qquad (4.36)$$

We thus arrive at (4.22), concluding the proof for Lemma 4.5. Theorem 4.4 then follows based on (4.16). □

The lower bound under the parametric setting given in Theorem 4.4 can be extended to the nonparametric setting following a similar line of proof. We have the following result (Agrawal, 1995 [4]).

Theorem 4.6 Lower Bound on Problem-Specific Regret in Nonparametric Setting: *Let $\mathcal{F}$ denote the set of possible arm distributions. Let π be a uniformly good policy over $\mathcal{F}$. We have,* *for all* $\mathbf{F} \in \mathcal{F}^N$,

$$\liminf_{T\to\infty} \frac{R_\pi(T;\mathbf{F})}{\log T} \geq \sum_{i:\mu_i<\mu_*} \frac{\mu_* - \mu_i}{D(F_i||(\mu_*)^+)}, \tag{4.37}$$

where, for a distribution $F \in \mathcal{F}$ *and a real number* c,

$$D(F||c^+) \triangleq \inf_{\mu>c} \inf_{F'\in\mathcal{F}:\mu_{F'}=\mu} D(F||F') \tag{4.38}$$

is the infimum of the KL divergence between F *and a distribution in* $\mathcal{F}$ *with a mean greater than* c.

4.2.2 THE MINIMAX REGRET

We now establish a lower bound on the regret under the minimax approach. This result traces back to as early as Vogel, 1960b [204]. The proof provided here largely follows that in Bubeck and Cesa-Bianchi, 2012 [45].

Theorem 4.7 Lower Bound on Minimax Regret:
Let $\mathcal{F}_b$ *denote the set of distributions with bounded support on* $[0, 1]$. *For every policy* π, *we have*

$$R_\pi(T;\mathcal{F}_b) \geq \frac{1}{20}\sqrt{NT}. \tag{4.39}$$

The minimax nature of the formulation implies that the above lower bound holds for all distribution sets that include bounded-support distributions, for example, sub-Gaussian, light-tailed, and heavy-tailed distributions. The question is in the achievability of this lower bound for these more general sets of distributions.

Techniques for establishing a lower bound on the minimax regret are quite different from that on the problem-specific regret. The intuition behind the order $O(\sqrt{NT})$ lower bound on the minimax regret is as follows. The Chernoff-Hoeffding inequality states that for any given confidence level, the sample mean calculated from n independent samples concentrates around the true mean in a confidence interval with length proportional to $\sqrt{1/n}$. For an N-armed bandit with a horizon length T, at least one arm is pulled no more than T/N times, resulting in a confidence interval of length at least in the order of $\sqrt{N/T}$. In other words, if the mean values of the best arm and the second best arm differ in the order of $\sqrt{N/T}$, these two arms cannot be differentiated with a diminishing probability of error. The resulting regret is in the order of $T\sqrt{N/T} = \sqrt{NT}$.

Proof. In establishing a lower bound on the minimax regret, it suffices to consider a specific distribution in the set of consideration. The choice here is the Bernoulli distribution $\mathcal{B}_p$ where p denotes the mean. Based on the intuition stated above, we consider a bandit problem in which one arm (the best arm) is $\mathcal{B}_{\frac{1}{2}+\epsilon}$ and the rest $N-1$ arms are $\mathcal{B}_{\frac{1}{2}}$. The worst-case regret of every policy is lower bounded by the regret incurred in this bandit problem. We then show that setting ϵ to be proportional to $\sqrt{N/T}$ maximizes the regret, thus leading to the tightest lower bound.

Let $\mathbf{F}_{(i)}$ denote the arm configuration under which arm i is $\mathcal{B}_{\frac{1}{2}+\epsilon}$ and the rest $N-1$ arms are $\mathcal{B}_{\frac{1}{2}}$. Let $\mathbb{E}_{(i)}$ and $\mathbb{P}_{(i)}$ denote, respectively, the expectation and the probability measure with respect to $\mathbf{F}_{(i)}$. We have, for an arbitrary policy π,

$$\begin{aligned}
\bar{R}_\pi(T) &\geq \max_{i=1,\ldots,N} R_\pi(T;\mathbf{F}_{(i)}) && (4.40)\\
&\geq \frac{1}{N}\sum_{i=1}^{N} R_\pi(T;\mathbf{F}_{(i)}) && (4.41)\\
&= \frac{\epsilon}{N}\sum_{i=1}^{N}\left(T-\mathbb{E}_{(i)}\left[\tau_i(T)\right]\right), && (4.42)
\end{aligned}$$

where $\tau_i(T)$ denotes the total number of times that arm i is pulled under π. In the following, we omit its dependence on T from the notation. The inequality (4.41) follows from the fact that the maximum is no smaller than the average, and the equality (4.42) follows from (4.6).

Next, we upper bound $\mathbb{E}_{(i)}\left[\tau_i\right]$ $(i=1,\ldots,N)$. The key idea is to use an arm configuration $\mathbf{F}_{(0)}$ with all arms following $\mathcal{B}_{\frac{1}{2}}$ as the benchmark and show that τ_i under $\mathbf{F}_{(i)}$ and $\mathbf{F}_{(0)}$ are sufficiently close in expectation.

Let $\mathbb{E}_{(0)}$, and $\mathbb{P}_{(0)}$ be the afore-defined notations under the benchmark configuration $\mathbf{F}_{(0)}$. Let $\mathbf{X}=(X_{\pi_1}(1),X_{\pi_2}(2),\ldots,X_{\pi_T}(T))$ denote the random reward sequence seen by policy π. For a deterministic policy π, the first arm selection π_1 is predetermined, and the arm selection π_t at time t is determined by the past seen rewards $X_{\pi_1}(1),\ldots,X_{\pi_{t-1}}(t-1)$. Thus, a given realization $\mathbf{x}=(x(1),\ldots,x(T))$ of the reward sequence determines the sequence of arm selection actions $(\pi_1,\ldots,\pi_T)$, hence the value of τ_i. The probability of $\mathbf{X}=\mathbf{x}$ under $\mathbf{F}_{(i)}$ is given by

$$\begin{aligned}
\mathbb{P}_{(i)}[\mathbf{X}=\mathbf{x}] &= \mathbb{P}_{(i)}[X_{\pi_1}(1)=x(1)]\\
&\quad\prod_{t=2}^{T}\mathbb{P}_{(i)}\left[X_{\pi_t}(t)=x(t)|X_{\pi_1}(1)=x(1),\ldots,X_{\pi_{t-1}}(t-1)=x(t-1)\right] && (4.43)\\
&= \prod_{t=1}^{T}\left(\mathcal{B}_{\frac{1}{2}}(x(t))\mathbb{I}[\pi_t\neq i]+\mathcal{B}_{\frac{1}{2}+\epsilon}(x(t))\mathbb{I}(\pi_t=i)\right). && (4.44)
\end{aligned}$$

Note that for a fixed reward sequence $\mathbf{x}$, the sequence of actions $(\pi_1,\ldots,\pi_T)$ under π is fixed. The indicator functions in (4.44) hence take fixed values.

Similarly, under $\mathbf{F}_{(0)}$, we have

$$\mathbb{P}_{(0)}(\mathbf{x})=\prod_{t=1}^{T}\mathcal{B}_{\frac{1}{2}}(x(t)). \tag{4.45}$$

We now bound the difference between $\mathbb{E}_{(i)}[\tau_i]$ and $\mathbb{E}_{(0)}[\tau_i]$ as follows. Let $\tau_i(\mathbf{x})$ denote the total number of plays on arm i under a given realization $\mathbf{x}$ of the reward sequence:

$$\begin{aligned}
\mathbb{E}_{(i)}[\tau_i] - \mathbb{E}_{(0)}[\tau_i] &= \sum_{\mathbf{x}} \tau_i(\mathbf{x}) \left(\mathbb{P}_{(i)}(\mathbf{x}) - \mathbb{P}_{(0)}(\mathbf{x})\right) && (4.46)\\
&\leq \sum_{\mathbf{x}:\mathbb{P}_{(i)}(\mathbf{x}) \geq \mathbb{P}_{(0)}(\mathbf{x})} \tau_i(\mathbf{x}) \left(\mathbb{P}_{(i)}(\mathbf{x}) - \mathbb{P}_{(0)}(\mathbf{x})\right) && (4.47)\\
&\leq T \sum_{\mathbf{x}:\mathbb{P}_{(i)}(\mathbf{x}) \geq \mathbb{P}_{(0)}(\mathbf{x})} \left(\mathbb{P}_{(i)}(\mathbf{x}) - \mathbb{P}_{(0)}(\mathbf{x})\right) && (4.48)\\
&\leq T||\mathbb{P}_{(i)} - \mathbb{P}_{(0)}||_{\text{TV}}, && (4.49)
\end{aligned}$$

where $||\mathbb{P}_{(i)} - \mathbb{P}_{(0)}||_{\text{TV}}$ denote the total variation distance between two distributions. Pinsker's inequality states that

$$||\mathbb{P}_{(i)} - \mathbb{P}_{(0)}||_{\text{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{P}_{(0)}||\mathbb{P}_{(i)})}. \qquad (4.50)$$

The KL divergence $D(\mathbb{P}_{(0)}||\mathbb{P}_{(i)})$, based on (4.45) and (4.44), can be written as

$$\begin{aligned}
D(\mathbb{P}_{(0)}||\mathbb{P}_{(i)}) &= \mathbb{E}_{(0)}\left[\log \frac{\mathbb{P}_{(0)}(\mathbf{X})}{\mathbb{P}_{(i)}(\mathbf{X})}\right] && (4.51)\\
&= \mathbb{E}_{(0)}\left[\sum_{t=1}^{T}\left(\log \frac{\mathcal{B}_{\frac{1}{2}}(X(t))}{\mathcal{B}_{\frac{1}{2}}(X(t))}\mathbb{I}(\pi_t \neq i) + \log \frac{\mathcal{B}_{\frac{1}{2}}(X(t))}{\mathcal{B}_{\frac{1}{2}+\epsilon}(X(t))}\mathbb{I}(\pi_t = i)\right)\right] && (4.52)\\
&= \sum_{t=1}^{T} \mathbb{E}_{(0)}\left[\log \frac{\mathcal{B}_{\frac{1}{2}}(X(t))}{\mathcal{B}_{\frac{1}{2}+\epsilon}(X(t))}\mathbb{I}(\pi_t = i)\right] && (4.53)\\
&= \frac{1}{2}\log \frac{1}{1-4\epsilon^2}\mathbb{E}_{(0)}[\tau_i]. && (4.54)
\end{aligned}$$

This leads to

$$\mathbb{E}_{(i)}[\tau_i] \leq \mathbb{E}_{(0)}[\tau_i] + \frac{T}{2}\sqrt{\log \frac{1}{1-4\epsilon^2}\mathbb{E}_{(0)}[\tau_i]}. \qquad (4.55)$$

Substituting this upper bound on $\mathbb{E}_{(i)}[\tau_i]$ into (4.42), we have

$$\begin{aligned}
\bar{R}_\pi(T) &\geq \frac{\epsilon}{N}\sum_{i=1}^{N}\left(T - \mathbb{E}_{(0)}[\tau_i] - \frac{T}{2}\sqrt{\log \frac{1}{1-4\epsilon^2}\mathbb{E}_{(0)}[\tau_i]}\right) && (4.56)\\
&= \epsilon T - \frac{\epsilon}{N}\sum_{i=1}^{N}\mathbb{E}_{(0)}[\tau_i] - \frac{\epsilon T}{2N}\sqrt{\log \frac{1}{1-4\epsilon^2}}\sum_{i=1}^{N}\sqrt{\mathbb{E}_{(0)}[\tau_i]} && (4.57)\\
&\geq \epsilon T - \frac{\epsilon}{N}T - \frac{\epsilon T}{2N}\sqrt{\log \frac{1}{1-4\epsilon^2}}\sqrt{NT}, && (4.58)
\end{aligned}$$

where the last line was obtained using the fact that $\sum_{i=1}^{N} \mathbb{E}_{(0)}[\tau_i] = T$ and $\sum_{i=1}^{N} \sqrt{\mathbb{E}_{(0)}[\tau_i]} \leq \sqrt{N \sum_{i=1}^{N} \mathbb{E}_{(0)}[\tau_i]}$ (the latter can be shown by applying Cauchy–Schwarz inequality to vectors $(1, 1, \ldots, 1)$ and $(\sqrt{\mathbb{E}_{(0)}[\tau_1]}, \ldots, \sqrt{\mathbb{E}_{(0)}[\tau_N]})$). We then arrive at the theorem by setting $\epsilon = \frac{1}{4}\sqrt{\frac{N}{T}}$ and using the inequality $\log \frac{1}{1-y} \leq 4 \log(\frac{4}{3}) y$ for $y \in [0, \frac{1}{4}]$.

For a randomized policy, the arm selection action at time t is random, drawn from a distribution determined by past reward observations. It can be equivalently viewed as following a deterministic policy $\pi = (\pi_1, \ldots, \pi_T)$ with a certain probability determined by the action distribution at each time in a product form. The lower bound thus applies to a randomized policy since it holds for each realization—a deterministic policy—of the randomized policy. □

4.3 ONLINE LEARNING ALGORITHMS

In this section, we present representative online learning algorithms that are asymptotically optimal or order optimal.

4.3.1 ASYMPTOTICALLY OPTIMAL POLICIES

Lai and Robbins developed a general method for constructing asymptotically optimal policies, i.e., policies that attain the asymptotic lower bound on the problem-specific regret given in Theorem 4.4.

Under the Lai–Robbins policy, two statistics are maintained for each arm. One is a point estimate $\hat{\mu}_i(t)$ of the mean of each arm i based on past observations. This statistic is used to ensure that an apparently superior arm is sufficiently exploited. The other statistic is the so-called upper confidence bound $U_i(t)$, which represents the *potential* of an arm: the less frequently an arm is played, the less confidence we have in its estimated mean, and the higher the potential of this arm. This statistic is used to ensure that each seemingly inferior arm is sufficiently explored to achieve a necessary level of confidence in the assertion on its inferiority. The upper confidence bound (UCB) of each arm depends on not only the number of plays on that arm but also the current time t in order to measure the *frequency* of exploration.

Based on these two statistics, the Lai–Robbins policy operates as follows. At each time t, the arm to be activated is chosen from two candidates: the leader $l(t)$ which is the arm with the largest estimated mean among all "well-sampled" arms defined as arms that have been played at least δt times for some predetermined constant $0 < \delta < \frac{1}{N}$, and the round-robin candidate $r(t)$ which rotates among all arms. The leader $l(t)$ is played if its estimated mean exceeds the UCB of the round-robin candidate $r(t)$. Otherwise, the round-robin candidate $r(t)$ is chosen. This policy is a modification of the follow-the-leader rule which greedily exploits without exploration.

What remains is to specify the point estimate $\hat{\mu}_i(t)$ of the mean and the UCB $U_i(t)$ for all $t = 1, 2, \ldots$ and $i = 1, \ldots, N$. Let $\tau_i(t)$ denote the number of times that arm i has been played

Algorithm 4.2 Lai–Robbins Policy

Notations: $\tau_i(t)$: the number of times that arm i is pulled in the first t plays;
$\hat{\mu}_i(t)$: the point estimate of the mean of arm i at time t;
$U_i(t)$: the upper confidence bound of arm i at time t;
Parameter: $\delta \in (0,\ 1/N)$.

```
Initialization: pull each arm once in the first $N$ rounds.
for $t = N + 1$ to $T$ do
    Identify the leader $l(t)$ among arms that have been pulled at least $(t-1)\delta$ times:
        $l_t \triangleq \arg \max_{i:\tau_i(t-1) \geq (t-1)\delta} \hat{\mu}_i(t-1).$
    Identify the round-robin candidate $r(t) = ((t-1) \bmod N) + 1$.
    if $\hat{\mu}_{l_t}(t-1) > U_{r_t}(t-1)$ then
        Pull arm $l(t)$.
    else
        Pull arm $r(t)$.
    end if
end for
```

up to and including t, and let $X_i(1), \ldots, X_i(\tau_i(t))$ denote the random rewards from these $\tau_i(t)$ plays on arm i. Let

$$\hat{\mu}_i(t) = h_{\tau_i(t)}(X_i(1), \ldots, X_i(\tau_i(t))) \tag{4.59}$$

$$U_i(t) = g_{t,\tau_i(t)}(X_i(1), \ldots, X_i(\tau_i(t))), \tag{4.60}$$

where $\{h_k\}_{k \geq 1}$ and $\{g_{t,k}\}_{t \geq 1, 1 \leq k \leq t}$ are measurable functions. Note that the point estimate and UCB may take different function forms for different sample size k and at different time t.

Lai and Robbins did not give an explicit construction of $\{h_k\}_{k \geq 1}$ and $\{g_{t,k}\}_{t \geq 1, 1 \leq k \leq t}$ for general reward distributions, but rather provided sufficient conditions on $\{h_k\}_{k \geq 1}$ and $\{g_{t,k}\}_{t \geq 1, 1 \leq k \leq t}$ for achieving asymptotic optimality.

Condition 4.8 *Lai–Robbins Condition on Mean Estimator and Upper Confidence Bound:*

(i) Strong consistency of the mean estimator for a well-sampled arm:

$$\mathbb{P}_\theta \left[\max_{\delta t \leq k \leq t} \left| h_k(X(1), \ldots, X(k)) - \mu(\theta) \right| > \epsilon \right] = o(t^{-1}) \tag{4.61}$$

for all $\theta \in \Theta$ and for all $\epsilon > 0$ and $0 < \delta < 1$.

(ii) UCB lower bounded by the mean estimate:

$$g_{t,k}((x(1),\ldots,x(k)) \geq h_k((x(1),\ldots,x(k)) \tag{4.62}$$

for all $t \geq k$ and for all $(x(1),\ldots,x(k))$.

(iii) UCB falling below the true mean with diminishing probability:

$$\mathbb{P}_\theta\left[g_{t,k}(X(1),\ldots,X(k)) \geq r \text{ for all } k \leq t\right] = 1 - o\left(t^{-1}\right) \tag{4.63}$$

for all $\theta \in \Theta$ and for all $r < \mu(\theta)$.

(iv) Monotonicity of UCB in t: for each fixed $k = 1, 2, \ldots$,

$$g_{t+1,k}(x(1),\ldots,x(k)) \geq g_{t,k}(x(1),\ldots,x(k)) \tag{4.64}$$

for all $t \geq k$ and for all $(x(1),\ldots,x(k))$.

(v) Asymptotically optimal exploration of suboptimal arms:

$$\lim_{\epsilon\downarrow 0}\left(\limsup_{t\to\infty}\frac{1}{\log t}\sum_{k=1}^{t}\mathbb{P}_\theta[g_{t,k}(X_(1),\ldots,X(k)) \geq \mu(\theta') - \epsilon]\right) \leq \frac{1}{D(\theta||\theta')}. \tag{4.65}$$

For all $\theta \in \Theta$ and for all $\theta' \in \Theta$ satisfying $\mu(\theta') > \mu(\theta)$.

The above conditions are relatively intuitive. In particular, Condition (i) ensures that the leader, determined by comparing estimated mean values among well-sampled arms, is indeed the best arm with high probability as the number of samples grows. For all distributions with a finite second moment, this condition is satisfied for the simple sample-mean estimator (see the proof of Theorem 1 of Chow and Lai, 1975 [63]):

$$h_k(X(1),\ldots,X(k)) = \frac{1}{k}\sum_{j=1}^{k} X(j). \tag{4.66}$$

Conditions (ii–iv) ensure that every arm is sufficiently explored by dictating that the UCB never falls below the true mean with high probability as time goes and the UCB of an arm kept passive grows with time. Condition (v) is a crucial condition to ensure that exploration of an inferior arm i with $\mu(\theta_i) < \mu(\theta^*)$ occurs no more than $(\log t)/D(\theta_i||\theta^*)$, asymptotically and in expectation, as dictated by the lower bound.

While the point estimator h_k of the mean can be set to the sample mean in most cases, the construction of the UCB $g_{t,k}$ satisfying conditions (ii)–(v) can be complex. Lai and Robbins, 1985 [126], gave constructions for four distributions: Bernoulli, Poisson, Gaussian, and Laplace. For the first two distributions, the mean value determines the entire distribution. For the last two, it is assumed that all arms have the same variance and it is known. For all four distributions, the point estimate of the mean is set to the maximum likelihood estimate (which is the sample mean for Bernoulli, Poisson, and Gaussian, and the sample median for Laplace), and the UCB is constructed based on generalized likelihood ratios. We illustrate the construction for the Bernoulli distribution in the example below.

Example 4.9 *Lai–Robbins Policy for Bernoulli Arms:*

Assume that arm i generates i.i.d. Bernoulli rewards with unknown parameter θ_i ($i = 1, \ldots, N$). We thus have $\mu(\theta) = \theta$ and $\Theta = (0, 1)$. The KL divergence $D(\theta||\lambda)$ is given by

$$D(\theta||\lambda) = \theta \log \frac{\theta}{\lambda} + (1 - \theta) \log \frac{1 - \theta}{1 - \lambda}. \tag{4.67}$$

It is easy to verify that the regularity conditions in Assumption 4.3 for the lower bound in Theorem 4.4 are satisfied.

At time t, the mean estimate $\hat{\mu}_i(t)$ and the UCB $U_i(t)$ of arm i are given by

$$\hat{\mu}_i(t) = \frac{1}{\tau_i(t)} \sum_{j=1}^{\tau_i(t)} X_i(j), \tag{4.68}$$

$$U_i(t) = \inf \left\{ \lambda \in (0, 1) : \lambda \geq \hat{\mu}_i(t) \text{ and } D(\hat{\mu}_i(t)||\lambda) \geq \frac{\log t}{\tau_i(t)} \right\}. \tag{4.69}$$

The above construction of the UCB $U_i(t)$ is based on the stochastic behaviors of the generalized log-likelihood ratio $\mathbb{P}_\theta[\sum_{j=1}^{k} \log \frac{f(x;\hat{\theta}_{j-1})}{f(x;\theta)}]$, where $\hat{\theta}_j = \hat{\theta}_j(X(1), \ldots, X(j))$ is an estimate of θ based on j random samples. It satisfies Conditions [C4.2–C4.5] for the Bernoulli as well as Gaussian and Poisson distributions as shown by Lai and Robbins, 1985 [126]. We point out that Lai and Robbins considered a general lower-bound sequence for $D(\hat{\mu}_i(t)||\lambda)$ that satisfies a set of conditions to ensure the asymptotic optimality of the resulting policy. To simplify the presentation and give an explicit implementation, we choose in (4.69) the sequence of $\frac{\log t}{\tau_i(t)}$, which can be easily shown to satisfy the conditions for the lower-bound sequence. While this choice ensures asymptotic optimality, a fine tuning may lead to better finite-time performance.

We omit the detailed derivation here. The intuition behind the above constructed UCB is that it is an overestimate of the mean as compared to the sample mean (the first condition in the right hand of (4.69)) and its divergence from the sample mean grows with $\frac{\log t}{\tau_i(t)}$, a measure of infrequency of exploration, to ensure asymptotically optimal exploration of a suboptimal arm.

The inexplicit characterization of $U_i(t)$ as an infimum can be worrisome when it comes to implementation. Fortunately, its explicit value is not necessary for implementing the Lai–Robbins policy. Since $D(\hat{\mu}_i(t)||\lambda)$ is a convex function in λ with minimum at $\lambda = \hat{\mu}_i(t)$, it is increasing in λ for $\lambda \geq \hat{\mu}_i(t)$. As a result, the condition $\hat{\mu}_{l_t}(t) \geq U_{r_t}(t)$ between the point estimate of the leader l_t and the UCB of the round-robin candidate r_t (Line 5 of Algorithm 4.2) is equivalent to

$$\hat{\mu}_{l_t}(t) \geq \hat{\mu}_{r_t}(t) \quad \text{and} \quad D(\hat{\mu}_{r_t}(t)||\hat{\mu}_{l_t}(t)) > \frac{\log t}{\tau_{r_t}(t)}, \tag{4.70}$$

which can be easily checked using only the mean estimates $\hat{\mu}_{l_t}(t)$ and $\hat{\mu}_{r_t}(t)$. Note that $D(\hat{\mu}_{r_t}(t)||\hat{\mu}_{l_t}(t))$ can be easily computed by plugging in the mean estimates into (4.67).

The construction of the UCB in the Lai-Robbins policy as given in (4.69) recently reappeared in the KL-UCB policy for Bernoulli arms (Garivier and Cappé, 2011 [84], Maillard, Munos, and Stoltz, 2011 [140], Cappé, et al., 2013 [54]). Specifically, the KL-UCB policy selects the following arm to play at time t:

$$a(t) = \arg \max_{i=1,\ldots,N} \sup \left\{ \lambda \in (0,1) : D(\hat{\mu}_i(t)||\lambda) \leq \frac{\log t}{\tau_i(t)} \right\}. \tag{4.71}$$

The equivalence of (4.71) to the construction of $U_i(t)$ in (4.69) can be easily seen from the fact that $D(\hat{\mu}_i(t)||\lambda)$ is increasing in λ for $\lambda \geq \hat{\mu}_i(t)$. The infimum over $\lambda \geq \hat{\mu}_i(t)$ satisfying $D(\hat{\mu}_i(t)||\lambda) \geq \frac{\log t}{\tau_i(t)}$ thus equals the supremum over λ satisfying $D(\hat{\mu}_i(t)||\lambda) \leq \frac{\log t}{\tau_i(t)}$. The only difference between these two policies is that the KL-UCB policy does not maintain a round-robin candidate. This modification was also considered by Lai in 1987 [125], who showed that the modified rule of selecting the arm with the highest UCB achieves asymptotic optimality for distributions in the exponential family. The order optimality of the KL-UCB policy for distributions with bounded support (other than Bernoulli) was also studied in [54, 84, 140].

Other asymptotically optimal policies include the sample-mean-based index policy by Agrawal, 1995 [4], for Bernoulli, Gaussian, Poisson, and exponential distributions. Except for Gaussian distribution, the indexes are only given implicitly and are difficult to compute. If the goal is only order optimality, these indexes can be constructed explicitly as discussed in the next section.

4.3.2 ORDER-OPTIMAL POLICIES

Asymptotically optimal policies achieve not only the optimal logarithmic regret order, but also the best leading constant as specified by the lower bounds as given in Theorems 4.4 and 4.6. Achieving asymptotic optimality requires a fine control on how often each suboptimal arm should be explored, which depends on the KL divergence between the reward distributions of a suboptimal arm and an optimal arm. If one is satisfied with order optimality, however, the problem is much more relaxed. Several simple and intuitive approaches suffice as discussed below.

Open-loop control of exploration-exploitation tradeoff A rather immediate solution is to exert an open-loop control on when to explore each of the N arms. Specifically, the time horizon is partitioned into two interleaving sequences: an exploration sequence and an exploitation sequence. At time instants belonging to the exploration sequence, each of the N arms has an equal chance to be selected, independent of past reward observations. This can be done either in a round-robin fashion or probabilistically by choosing an arm uniformly at random. At time instants in the exploitation sequence, the arm with the greatest sample mean (or a properly chosen point estimator of the mean) is played. This approach separates in time the two objectives of exploration and exploitation, and the tradeoff between these two is reflected in the cardinality of the exploration sequence (equivalently, the cardinality of the exploitation sequence).

The partition of the time horizon into exploration and exploitation sequences can be done deterministically or probabilistically. In a deterministic partition, the sequence of time instants for exploration is explicitly specified. This approach was first suggested by Robbins, 1952 [168], when the frequentist version of the bandit problems was first posed, and later studied in detail by Agrawal and Teneketzis, 1989 [8], Yakowitz and Lowe, 1991 [215], and Vakili, Liu, and Zhao, 2013 [194]. This approach was referred to as Deterministic Sequencing of Exploration and Exploitation (DSEE) in [194]. In a probabilistic partition, each time instant t is assigned to the exploration or the exploitation sequence with probability ϵ and $1-\epsilon$, respectively. This approach is known as the ϵ-greedy policy (see Sutton and Barto, 1998 [187] and Auer, Cesa-Bianchi, and Fischer, 2002 [23]).

What remains to be specified is the cardinality of the exploration sequence. On the one hand, the regret order is lower bounded by the cardinality of the exploration sequence since a fixed fraction of the exploration sequence is spent on suboptimal arms. On the other hand, the exploration sequence needs to be chosen sufficiently dense to ensure effective learning of the best arm. The key issue here is to find the minimum cardinality of the exploration sequence that ensures a reward loss in the exploitation sequence caused by an incorrectly identified best arm having an order no greater than the cardinality of the exploration sequence. The lower bound established by Lai and Robbins (Theorem 4.4) shows that a consistent policy must explore each arm at least in the order of $\log T$ times. This turns out to be also sufficient for policies with an open-loop control of the exploration-exploitation tradeoff, leading to the optimal logarithmic regret order. For the ϵ-greedy policy, setting ϵ as a function of time t that diminishes to zero at a rate of $1/t$ gives a (random) exploration sequence with an (expected) cardinality in the order of $\log T$.

Compared with Lai and Robbins policy, ϵ-greedy and DSEE are much simpler and more general due to the nonparametric setting. This comparison highlights the often formidable gap between achieving asymptotic optimality and merely order optimality. To achieve asymptotic optimality, the exploration of each suboptimal arm i needs to be finely controlled to ensure not only the order of the exploration is $\log T$, but also the leading constant converges to $1/D(\theta_i||\theta^*)$. For order optimality, however, it is not necessary to differentiate suboptimal arms in exploration, and it suffices to explore all arms equally often, provided that the order of the exploration time is logarithmic with a leading constant sufficiently large. Setting the leading constant in the cardinality of the exploration sequence often requires a lower bound on the mean difference between the best and the second best arms. Setting it to an arbitrarily slowly diverging sequence circumvent this issue when such prior information is unavailable (see details in [194]).

Index-type policies The computation of the Lai–Robbins policy can be complex for general reward distributions, requiring the entire sequence of past rewards received from each arm. This motivated Agrawal, 1995 [4], to develop index-type policies that use only certain simple statistics of reward sequences. Two key statistics to prioritize arms are the sample mean $\hat{\mu}_i(t)$ calculated from past observations and the number $\tau_i(t)$ of times that arm i has been played in the past t plays. The larger $\hat{\mu}_i(t)$ is or the smaller $\tau_i(t)$ is, the higher the priority given to this arm in arm selection. The tradeoff between exploration and exploitation is reflected in how these two statistics are combined together for arm selection.

Agrawal, 1995 [4] developed order-optimal index policies for several distribution types, in which an index $I_i(t+1)$ is computed at time $t+1$ (i.e., after t plays) for each arm and the arm with the largest index is chosen. The index has the following forms:

$$I_i(t+1) = \begin{cases} \hat{\mu}_i(t) + (2\frac{\log t + 2\log\log t}{\tau_i(t)})^{1/2}, & \text{Gaussian} \\ \hat{\mu}_i(t) + \min\left\{\frac{1}{2}\sqrt{\frac{2\log t + 4\log\log t}{\tau_i(t)}},\ 1\right\}, & \text{Bernoulli} \\ \hat{\mu}_i(t) + \min\left\{\frac{1}{2}\sqrt{\frac{2a\log t + 4a\log\log t}{\tau_i(t)}},\ a\right\}, & \text{Poisson} \\ \hat{\mu}_i(t) + b\min\left\{\sqrt{\frac{2\log t + 4\log\log t}{\tau_i(t)}},\ 1\right\}, & \text{Exponential} \end{cases}, \tag{4.72}$$

where constants a and b are, respectively, upper bounds on possible parameter values in the corresponding families of Poisson and exponential reward distributions. The index for Gaussian distributions (with known variance) achieves asymptotic optimality.

These index forms for different distribution types have the same structure: the sample mean plus a UCB term determined by $\tau_i(t)$. Specifically, the second term of the index ensures that each arm is explored in the order of $\log t$ times as dictated by the lower bound. For an arm sampled in an order smaller than $\log t$, its index, dominated by the second term, will be sufficient large for large t to ensure further exploration.

Auer, Cesa-Bianchi, and Fischer, 2002 [23], modified Agrawal's sample-mean-based index policies to offer order optimality for the family $\mathcal{F}_b$ of distributions with bounded support (normalized to $[0, 1]$). This is the much celebrated UCB-α policy with the following index form similar to those in (4.72):

$$I_i(t+1) = \hat{\mu}_i(t) + \sqrt{\frac{\alpha \log t}{\tau_i(t)}}, \tag{4.73}$$

where $\alpha > 1$ is a policy parameter fixed to 2 in the UCB1 policy in Auer, Cesa-Bianchi, and Fischer, 2002 [23].

Algorithm 4.3 The UCB-α Policy

Notations: $\tau_i(t)$: the number of times that arm i is pulled in the first t plays;
$\hat{\mu}_i(t)$: the sample mean of the $\tau_i(t)$ rewards from arm i.
Parameters: $\alpha > 1$.

```
Initialization: pull each arm once in the first $N$ rounds.
for $t = N + 1$ to $T$ do
    Compute the index $I_i(t)$ of each arm $i$ as given in (4.73)
    Pull the arm $a(t)$ with the greatest index and observe reward $x_{a(t)}$.
    Update the sample mean and the number of plays for arm $a(t)$:
        $\tau_{a(t)}(t) = \tau_{a(t)}(t-1) + 1$
        $\hat{\mu}_{a(t)}(t) = \frac{1}{\tau_{a(t)}(t)} \left(\hat{\mu}_{a(t)}(t-1)\tau_{a(t)}(t-1) + x_{a(t)}\right)$
end for
```

The work by Auer, Cesa-Bianchi, and Fischer, 2002 [23] gives an easily accessible regret analysis that establishes not only the regret order but also a finite-time regret bound for all T. We present this result in the theorem below.

Theorem 4.10 *Assume that the random reward of each arm takes value in* $[0, 1]$. *Define*

$$\Delta_i \triangleq \mu_* - \mu_i. \tag{4.74}$$

Then for every horizon length T, the problem-specific regret of UCB-α with $\alpha > 1$ is upper bounded by

$$R_{UCB\text{-}\alpha}(T; \mathbf{F}) \leq \sum_{i:\Delta_i > 0} \left(\frac{4\alpha}{\Delta_i} \log T + \frac{1}{\alpha - 1} \Delta_i \right). \tag{4.75}$$

Proof. We bound the regret by bounding $\tau_i(T)$ for each suboptimal arm i (i.e., an arm i with $\Delta_i > 0$). In the following, all quantities associated with an optimal arm are indicated with a subscript $*$. Let π denote the UCB-α policy. By definition, we have

$$\tau_i(T) = 1 + \sum_{t=N+1}^{T} \mathbb{I}(\pi(t) = i). \tag{4.76}$$

The event $\pi(t) = i$ implies that $I_i(t) \geq I_*(t)$ (although not vice versa). To have $I_i(t) \geq I_*(t)$, at least one of the following must hold:

$$\begin{aligned} \Delta_i &< 2\sqrt{\alpha \frac{\log t}{\tau_i(t)}}, && (4.77) \\ I_*(t) &\leq \mu_*, && (4.78) \\ I_i(t) &\geq \mu_i + 2\sqrt{\alpha \frac{\log t}{\tau_i(t)}}. && (4.79) \end{aligned}$$

This can be seen by noticing that if none of the above three inequalities is true, then the gap Δ_i between μ_* and μ_i is at least $2\sqrt{\alpha \frac{\log t}{\tau_i(t)}}$ and the index $I_*(t)$ of the optimal arm exceeds μ_* while the index $I_i(t)$ of arm i does not make up the minimum gap between μ_* and μ_i, leading to $I_i(t) < I_*(t)$.

Set $t_0 = \frac{4\alpha \log T}{\Delta_i^2}$. We readily have $t_0 \geq \frac{4\alpha \log t}{\Delta_i^2}$ for $t = 1, \ldots, T$. For $\tau_i(t) > t_0$, (4.77) is false, thus at least one of (4.78) and (4.79) must be true. We then have

$$\tau_i(T) \leq t_0 + \sum_{t=t_0+1}^{T} \mathbb{I}\left[(4.78) \text{ or } (4.79) \text{ holds}\right], \tag{4.80}$$

where the inequality follows from the simple fact that after arm i is pulled t_0 times, the decision time t is at least $t_0 + 1$. Taking expectation and applying the union bound, we have

$$\mathbb{E}[\tau_i(T)] \leq t_0 + \sum_{t=t_0+1}^{T} \mathbb{P}\left[I_*(t) \leq \mu_*\right] + \sum_{t=t_0+1}^{T} \mathbb{P}\left[I_i(t) \geq \mu_i + 2\sqrt{\alpha \frac{\log t}{\tau_i(t)}}\right]. \tag{4.81}$$

Next, we use the Hoeffding inequality to bound $\mathbb{P}\left[I_*(t) \leq \mu_*\right]$ and $\mathbb{P}\left[I_i(t) \geq \mu_i + 2\sqrt{\alpha \frac{\log t}{\tau_i(t)}}\right]$. Let $\hat{\mu}_{*,s}$ denote the sample mean of the optimal arm calculated from s samples:

$$\begin{aligned}
\mathbb{P}\left[I_*(t) \leq \mu_*\right] &= \mathbb{P}\left[\hat{\mu}_*(t) + \sqrt{\alpha \frac{\log t}{\tau_*(t)}} \leq \mu_*\right] && (4.82)\\
&\leq \mathbb{P}\left[\exists s \in \{1, \ldots, t\} : \hat{\mu}_{*,s} + \sqrt{\alpha \frac{\log t}{s}} \leq \mu_*\right] && (4.83)\\
&\leq \sum_{s=1}^{t} \mathbb{P}\left[\hat{\mu}_{*,s} + \sqrt{\alpha \frac{\log t}{s}} \leq \mu_*\right] && (4.84)\\
&= \sum_{s=1}^{t} \mathbb{P}\left[\hat{\mu}_{*,s} - \mu_* \leq -\sqrt{\alpha \frac{\log t}{s}}\right] && (4.85)\\
&\leq \sum_{s=1}^{t} \frac{1}{t^{2\alpha}} && (4.86)\\
&= \frac{1}{t^{2\alpha-1}}, && (4.87)
\end{aligned}$$

where (4.83) follows by considering all possible values of $\tau_*(t)$ and (4.86) follows from the following Chernoff–Hoeffding inequality.

Lemma 4.11 Chernoff–Hoeffding Bound:
Let $X(1), \ldots, X(s)$ *be random variables with common support of* $[0, 1]$. *Suppose that* $\mathbb{E}[X(j)|X(1), \ldots, X(j-1)] = \mu$ *for* $j = 1, \ldots, s$. *Let* $\hat{\mu}_s = \frac{1}{s}\sum_{j=1}^{s} X(j)$. *Then for all* $a \geq 0$,

$$\mathbb{P}[\hat{\mu}_s - \mu \geq a] \leq e^{-2a^2 n} \quad \textit{and} \quad \mathbb{P}[\hat{\mu}_s - \mu \leq -a] \leq e^{-2a^2 n}. \tag{4.88}$$

Similarly,

$$\mathbb{P}\left[I_i(t) \geq \mu_i + 2\sqrt{\alpha \frac{\log t}{\tau_i(t)}}\right] = \mathbb{P}\left[\hat{\mu}_i(t) - \mu_i \geq -\sqrt{\alpha \frac{\log t}{\tau_i(t)}}\right] \leq \frac{1}{t^{2\alpha-1}}. \tag{4.89}$$

We then have

$$\begin{aligned}
\mathbb{E}[\tau_i(T)] &\leq t_0 + \sum_{t=t_0+1}^{T} \frac{2}{t^{2\alpha-1}} && (4.90)\\
&\leq t_0 + \int_1^{\infty} \frac{2}{t^{2\alpha-1}} \mathrm{d}t && (4.91)\\
&= t_0 + \frac{1}{\alpha - 1}. && (4.92)
\end{aligned}$$

The theorem then follows by plugging in the above bound on $\mathbb{E}[\tau_i(T)]$ to (4.6).

□

The worst-case regret of the UCB-α policy is in the order of $O(\sqrt{NT \log T})$, which does not match with the lower bound on minimax regret given in Theorem 4.7. Audibert and Bubeck, 2009 [19] proposed a modification of the UCB-α policy and showed that the modified policy, referred to as MOSS (Minimax Optimal Strategy in the Stochastic case), achieves the optimal $\sqrt{NT}$ minimax regret order while preserving the $\log T$ problem-specific regret order. The modified index has the following form:

$$I_i(t+1) = \hat{\mu}_i(t) + \sqrt{\frac{\max\{\log \frac{T}{N\tau_i(t)}, 0\}}{\tau_i(t)}}. \tag{4.93}$$

The UCB-α policy and its variants are often presented and analyzed for reward distributions with bounded support. Extensions to more general distribution families such as sub-Gaussian and locally sub-Gaussian (i.e., light-tailed) distributions are relatively straightforward; see treatments in Bubeck and Cesa-Bianchi, 2012 [45] and Lattimore and Szepesvari, 2019 [129]. Dealing with heavy-tailed distributions is more involved; see the extension of the UCB-α policy to heavy-tailed distributions by Bubeck, Cesa-Bianchi, and Lugosi, 2013 [50], and that of the DSEE policy in Vakili, Liu, Zhao, 2013 [194].

4.4 CONNECTIONS BETWEEN BAYESIAN AND FREQUENTIST BANDIT MODELS

While the Bayesian and frequentist frameworks lead to different bandit models with different optimality criteria, learning algorithms developed within one framework can be applied and evaluated in the other. This also brings the question on the existence of learning algorithms that are optimal in some sense from both the Bayesian and the frequentist viewpoints. We see in this subsection that the answer is positive.

4.4.1 FREQUENTIST APPROACHES TO BAYESIAN BANDITS

Consider first the direction of taking frequentist approaches (e.g., the UCB-type policies) as solutions to the Bayesian bandit problem. If we restrict ourselves to the original bandit problem of sampling unknown processes (referred to as the bandit sampling processes) where the arm states are informational and state transitions obey Bayes rule, the applicability of frequentist approaches can be readily seen. We simply ignore the prior probabilistic knowledge on the arm parameters and directly apply the frequentist algorithms. Obviating the need to update posterior distributions and using simple statistics such as the sample mean and the number of plays, the resulting policies have an edge in computational efficiency. The main question is the performance measured under the Bayesian criteria.

If the objective is the total discounted reward, ignoring prior knowledge and the discounting factor are likely to incur significant performance loss. This is due to the fact that discounting effectively shortens the decision horizon (hence amplifying the impact of prior knowledge on the overall performance) and places heavier weights on decisions at the beginning of the time horizon. If, however, the objective is the average reward over an infinite horizon, prior knowledge has diminishing effect on the overall performance. All frequentist algorithms with a sublinear regret order, not even necessarily the optimal sublinear order, offer the maximum average reward μ_*. This echoes the statement in Section 3.5.2 that the average-reward criterion is unselective. The connection between asymptotic/order-optimality in the frequentist regret measure and the optimality under the Bayesian gain-bias and sensitive-discount criteria as discussed in Section 3.5.2 is worth exploring.

A performance measure that bridges the Bayesian and frequentist formulations is the *Bayesaian regret* introduced by Lai, 1987 [125], which is the averaged problem-specific regret over the problem instances with a given prior distribution. This corresponds to the Bayes risk commonly used by the Bayesian school. Consider the parametric setting discussed at the beginning of Section 4.2.1. The Bayesian regret of a policy π is given by

$$\int_{\boldsymbol{\theta}} R_\pi(T;\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{4.94}$$

where $R_\pi(T;\boldsymbol{\theta})$ is the problem-specific regret in (4.16) and $p(\boldsymbol{\theta})$ the prior distribution of the arm parameters $\boldsymbol{\theta}$. Bayesian regret is not a function of $\boldsymbol{\theta}$, but a function of the prior $p(\boldsymbol{\theta})$.

A significant result in Lai, 1987 [125] is that a modified version of the Lai–Robbins policy that pulls the arm with the highest UCB is asymptotically optimal from both the Bayesian and frequentist viewpoints for distributions in the exponential family. Specifically, this policy achieves, with the corresponding optimal leading constants, both the optimal $\log T$ problem-specific regret order and the optimal $\log^2 T$ Bayesian regret order under sufficiently regular prior distributions.

The Bayesian regret is considerably harder to analyze than its frequentist counterparts. To bound the asymptotic Bayesian regret of a policy, we need to establish asymptotic properties of its problem-specific regret that hold *uniformly* over a range of $\boldsymbol{\theta}$ in the parameter space so that the integral in (4.94) can be evaluated. The wider this range of uniformity, the broader the class of prior distributions under which the Bayesian regret can be bounded. An alternative approach is to consider the worst-case prior specific to the policy and the horizon length. This brings us to the minimax setting, with the same optimal order of $\sqrt{NT}$ as in the frequentist framework.

4.4.2 BAYESIAN APPROACHES TO FREQUENTIST BANDITS

Bayesian approaches can be used to solve a frequentist bandit problem by adopting a fictitious prior distribution of the arm parameters. In a nonparametric setting, the fictitious prior can be set on the unknown mean μ_i of each arm, and in addition, a fictitious likelihood function

$\tilde{f}_i(x; \mu_i)$ is needed for calculating the posterior distribution of μ_i based on observations of the reward X_i.

There are two considerations in choosing the fictitious prior distributions and likelihood functions: the resulting frequentist regret performance and the computation involved in the update of the posterior distributions. As has been shown, the frequentist regret performance, particularly in terms of achieving asymptotic/order optimality, depends on certain matching between the assumed prior/likelihood function with the underlying arm reward distributions. In terms of the computation involved in posterior updates, it is desirable to choose a prior and a likelihood function that form a conjugate pair, that is, the posterior calculated from the prior and the likelihood function belongs to the same probability distribution family as the prior. Examples are Beta prior for Bernoulli rewards and Gaussian prior (on the mean) for Gaussian rewards as detailed below.

We use Thompson Sampling as an example to illustrate the idea of employing Bayesian algorithms to Frequentist bandit models. As detailed in Example 2.3, Thompson Sampling is a randomized policy that selects each arm with a probability equal to the posterior probability of it being the optimal one. It is a suboptimal solution to the Bayesian bandit model under both infinite horizon with discounted-reward criterion and the finite horizon problem as considered by Thompson in his original work of 1933 [190]. If we let the horizon length T tend to infinity, however, Thompson Sampling has recently been shown to be asymptotically optimal in terms of the problem-specific regret for several frequentist bandit problems.

Consider first the parametric setting where the reward distribution of arm i is $f(x; \theta_i)$ (viewed as a function of x parameterized by θ_i). It is also the likelihood function when viewed as a function of the unknown parameter θ_i for a given observation x. With the true likelihood function known, we only need to assume a fictitious prior distribution $\tilde{p}_{\theta_i}$ for each arm. An implementation of Thompson Sampling is given in Algorithm 4.4, which generalizes Example 2.3 that considered a uniform prior and Bernoulli reward distributions.

For a nonparametric setting, consider the class $\mathcal{F}_b$ of distributions with support over [0, 1]. The mean μ_i of each arm is set to be the unknown parameter for which a prior will be assumed and the posterior distribution will be updated based on a fictitious likelihood function. We present below two ways of employing Thompson Sampling in this setting developed by Agrawal and Goyal, 2012, 2013 [6, 7].

The first one is to use the beta distribution Beta(α, β) as the prior and assume a Bernoulli reward distribution with mean μ_i for arm i $(i = 1, \ldots, N)$. A Bernoulli distributed reward observation X gives the following likelihood functions of μ_i:

$$f(X = 1; \mu_i) = \mu_i, \quad f(X = 0; \mu_i) = 1 - \mu_i. \tag{4.95}$$

Being the conjugate prior for the likelihood functions induced by the Bernoulli observations, a Beta(α, β) prior, upon observing a Bernoulli reward X, leads to a posterior that is simply Beta$(\alpha + 1, \beta)$ if $X = 1$ and Beta$(\alpha, \beta + 1)$ if $X = 0$. The posterior in each decision period thus remains in the beta family with the two parameters α and β simply counting the number of

Algorithm 4.4 Thompson Sampling for Parametric Frequentist Bandits

```
Input: $\tilde{p}_{\theta_i}^{(1)}$: a fictitious prior of $\theta_i$ $(i = 1, \dots, N)$;

for $t = 1$ to $T$ do
    Generate a random sample $\tilde{\theta}_i(t)$ from $\tilde{p}_{\theta_i}^{(t)}$ $(i = 1, \dots, N)$.
    Compute, for $i = 1, \dots, N$,
        $\mu(\tilde{\theta}_i(t)) = \int_x x f\left(x; \tilde{\theta}_i(t)\right) \mathrm{d}x.$
    Play arm $\tilde{a} = \arg\max_{i=1,\dots,N} \mu(\tilde{\theta}_i(t))$ and receive reward $x_{\tilde{a}}$.
    Update the posterior distribution for arm $\tilde{a}$: for all $z \in \Theta$,
        $\tilde{p}_{\theta_{\tilde{a}}}^{(t+1)}(z) = \frac{\tilde{p}_{\theta_{\tilde{a}}}^{(t)}(z) f(x_{\tilde{a}}; z)}{\int_y f(x_{\tilde{a}}; y) \tilde{p}_{\theta_{\tilde{a}}}^{(t)}(y) \mathrm{d}y}$
end for
```

successes and failures. The initial prior can be set to Beta$(1, 1)$, which is the uniform distribution over $[0, 1]$.

What remains to be specified is how to used the true reward observations with support $[0, 1]$ in the above likelihood functions in (4.95) that assumes Bernoulli rewards of 0 or 1. A method introduced in Agrawal and Goyal, 2012 [6], is to, upon receiving a reward $x \in [0, 1]$, toss a coin with bias x and use the outcome as the input to the (fictitious) likelihood functions. This translation of a general reward distribution over $[0, 1]$ to a Bernoulli distribution preserves the regret bounds developed under Bernoulli rewards.

The second approach to using Thompson Sampling in the nonparametric frequentist setting is to adopt a Gaussian prior and assume a Gaussian reward distribution with mean μ_i and unit variance for arm i $(i = 1, \dots, N)$. Since the Gaussian family is conjugate to itself, the posterior distribution is Gaussian at each time. The initial prior can be set to the Gaussian distribution $\mathcal{N}(0, 1)$. The posterior distribution at the beginning of decision time $t > 1$ for arm i is then $\mathcal{N}(\tilde{\mu}_i(t-1), \frac{1}{\tau_i(t-1)+1})$, where $\tau_i(t-1)$, as usual, is the number of plays on arm i in the past $t-1$ plays and $\tilde{\mu}_i(t-1) = \frac{1}{\tau_i(t-1)+1} \sum_{j=1}^{\tau_i(t-1)} x_i(j)$ with $x_i(1), \dots, x_i(\tau_i(t-1))$ denoting the $\tau_i(t-1)$ reward observations from arm i. A random sample is then drawn from the posterior distribution of each arm, and the arm with the largest sample value is played at time t. Note that the true rewards are within the support of the assumed Gaussian distribution. The observed rewards are directly used in the posterior updates.

In terms of the problem-specific regret, the asymptotic optimality of Thompson Sampling with Beta prior for Bernoulli arms has been established by Agrawal and Goyal, 2012 [6] and Kaufmann, Korda, and Munos, 2012 [110]. Its asymptotic optimality under properly chosen priors has been shown for single-parameter exponential family by Korda, Kaufmann, and Munos, 2013 [120], and for Gaussian arms by Honda and Takemura, 2014 [102]. The nearly optimal worst-case regret of Thompson Sampling for arms with bounded support has been established by Agrawal and Goyal, 2013 [7]. Strong empirical performance of Thompson Sampling for frequentist bandits is also well documented.

Another Bayesian approach to frequentist bandits is the Bayes–UCB algorithm proposed and analyzed by Kaufmann, Cappé, and Garivier, 2012 [109] and Kaufmann, 2018 [108]. In Bayes–UCB, the quantiles of the posterior distribution, in the role of the UCB-based index, is used for arm selection. The asymptotic optimality of Bayes–UCB in terms of the problem-specific regret has been shown for Bernoulli arms (Kaufmann, Cappé, and Garivier, 2012 [109]) and single-parameter exponential families (Kaufmann, 2018 [108]).

CHAPTER 5

Variants of the Frequentist Bandit Model

The set of variants of the frequentist bandit model that have been considered in the literature is much richer than that of the Bayesian bandit model. This may partly due to the elasticity of the order-optimality objective widely accepted for frequentist bandit problems. Assumptions can be relaxed, additional model complexities can be introduced, yet the problem remains analytically tractable under the objective of order optimality. For the Bayesian bandit model, the objective of optimizing the total discounted reward is unforgiving of any suboptimal action during the entire course of the decision process. As we have seen in Chapter 3, the optimality of the Gittins index policy can be fragile, as is often the case with optimal policies for MDP problems with strikingly simple and elegant analytical structures.

We cover in this chapter major first-order variants that focus on a single aspect of the defining features of the bandit model. Higher-order variants can be easily constructed, if such studies are of interest both intellectually and application-wise. Given the vast landscape of the existing literature and the fast pace of progress in this active research area, the coverage of this chapter, in terms of its inclusion of existing studies, old and new, as well as technical details involved, is far from comprehensive. Our focus is on categorizing various formulations of these bandit variants, outlining general ideas of solution methods, and providing references to representative studies. Rather than providing a detailed presentation of specific algorithms and analysis, we hope to highlight the connections and differences between these variants and the canonical model, and the new challenges and how they might be addressed.

5.1 VARIATIONS IN THE REWARD MODEL

In the canonical frequentist bandit model, each arm i $(i = 1, \ldots, N)$ is associated with an i.i.d. reward process $\{X_i(t)\}_{t\geq 1}$ drawn from a fixed unknown distribution $F_i(x)$. In this section, we consider bandit variants that adopt more general models for the arm reward processes.

One direction is to allow temporal dependencies in the reward processes. A natural candidate is the Markov process, leading to frequentist bandit models with intricate connections with the Bayesian bandit models discussed in Chapters 2 and 3. In particular, depending on whether the arm reward state continues to evolve when the arm is passive, two variants, referred to as the *rested frequentist bandit* and the *restless frequentist bandit*, have been considered.

The second direction is to consider nonstationary reward processes where the unknown distribution $F_i(x)$ of each arm is time varying. Three different temporal-variation models have been considered: the abrupt-change model, the continuous-drift model, and the total-variation model.

The third direction, which represents a much more significant departure from the canonical model, is to forgo the stochastic nature of the reward models and consider *deterministic* reward sequences potentially designed by an adversary that reacts to the player's strategy. Referred to as the nonstochastic bandit or the adversarial bandit, this variant is receiving increasing attention and demands a book of its own. We give here only a brief discussion.

5.1.1 RESTED MARKOV REWARD PROCESSES

Consider a bandit model where random rewards from successive plays of arm i form a finite-state, irreducible, and aperiodic Markov chain with state space $\mathcal{S}_i$ and an unknown transition probability matrix $\mathbf{P}_i$. When arm i is not engaged, its reward state remains frozen. This gives us a bandit model with *rested* Markov reward processes.

A proxy for oracle and regret decomposition: The first step is to characterize the optimal strategy of the oracle who has the knowledge of the transition probability matrix $\mathbf{P}_i$ of each arm. This is the Bayesian bandit model under the finite-horizon total-reward criterion. As discussed in Section 3.1.3, the optimal policy for this finite-horizon problem is in general nonstationary and analytically intractable. The lack of an analytical characterization of the benchmark presents a major obstacle to regret analysis and policy development.

With an objective focusing on the order of the regret as T approaches infinity, however, it suffices to consider a proxy for the oracle that offers the same order of performance as the oracle but better analytical tractability. This approach of finding a proxy benchmark turns out to be quite powerful and will be called upon multiple times in this chapter.

For the bandit model with rested reward processes, the single-arm policy that always plays the arm with the maximum limiting reward rate suffices as a proxy benchmark. Specifically, let $\{\gamma_i(x)\}_{x\in\mathcal{S}_i}$ denote the unique stationary distribution of arm i. The asymptotic average reward per play offered by arm i is thus given by, for any initial state $x_0 \in \mathcal{S}_i$,

$$\mu_i \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} X_i(t) \mid X_i(1) = x_0\right] = \sum_{x\in\mathcal{S}_i} x\gamma_i(x), \tag{5.1}$$

where we have used the fact that a finite-state, irreducible, and aperiodic Markov chain is ergodic with a limiting distribution given by the unique stationary distribution. Due to the rested nature and the ergodicity of the reward processes, the optimal oracle strategy that exploits the initial transient behaviors of arms will eventually settle down on the optimal arm $i_* = \arg\max_{i=1,\dots,N} \mu_i$ after a finite number of arm switchings (bounded by the mixing times of the Markov reward processes). Consequently, the performance gap between the single-arm policy that always plays arm i_* and the optimal oracle strategy is upper bounded by a constant independent of the initial arm reward states and the horizon length T. Thus, using $\mu_* T$ as the proxy benchmark preserves the asymptotic behavior of regret.

By applying Wald's identity to the regenerative cycles of each Markov reward process, the (proxy) regret of policy π can be written as

$$R_\pi(T; \{\mathbf{P}_i\}) = \sum_{i:\mu_i < \mu_*} (\mu_* - \mu_i)\mathbb{E}_\pi[\tau_i(T)], \tag{5.2}$$

which has the same expression as (4.6) for i.i.d. rewards, except that μ_i here is determined by the limiting distribution of the corresponding Markov chain. In analyzing the regret performance, the key quantities are again the expected times $\mathbb{E}_\pi[\tau_i(T)]$ spent on each suboptimal arm. As discussed below, the lower bound results and learning policies for the canonical frequentist bandit model can be readily extended by employing large deviation results on the empirical distribution and the empirical transition-count matrix of finite-state Markov chains.

Regret lower bound: Existing results on bandits with rested Markovian reward processes mainly focus on the uniform-dominance approach. The first result was given by Anantharam, Varaiya, and Walrand, 1987 [16], adopting a similar univariate parametric model as in the work by Lai and Robbins, 1985 [126], for i.i.d. reward processes. Specifically, all arms have a common state space $\mathcal{S}$, and the transition probability matrix $\mathbf{P}(\theta_i)$ of arm i is determined by an unknown parameter θ_i belonging to a parameter space Θ. It is assumed that for all reward states $x, x' \in \mathcal{S}$ and for all parameter values $\theta, \theta' \in \Theta$, if $\mathbf{P}_{x,x'}(\theta) > 0$, then $\mathbf{P}_{x,x'}(\theta') > 0$. Furthermore, $\mathbf{P}(\theta)$ is irreducible and aperiodic for all $\theta \in \Theta$. The initial state distribution is assumed to be positive for all states.

The problem-specific lower bound echoes that of Lai and Robbins, 1985 [126]. Specifically, it was shown that under similar regularity conditions (specifically, the continuity of $D(\theta||\lambda)$ in $\lambda > \theta$ for each fixed $\theta \in \Theta$ and the denseness of Θ), for every uniformly good policy π and every arm parameter set $\boldsymbol{\theta}$ such that $\mu(\theta_i)$ are not all equal, we have

$$\liminf_{T\to\infty} \frac{R_\pi(T; \boldsymbol{\theta})}{\log T} \geq \sum_{i:\mu(\theta_i)<\mu_*} \frac{\mu_* - \mu(\theta_i)}{D(\theta_i||\theta_*)}, \tag{5.3}$$

where

$$D(\theta_i||\theta_*) \triangleq D(\mathbf{P}(\theta_i)||\mathbf{P}(\theta_*)) \tag{5.4}$$

is the KL divergence between two stochastic matrices. Recall that for two irreducible and aperiodic stochastic matrices $\mathbf{P}$ and $\mathbf{Q}$ satisfying $\mathbf{P}_{x,x'} > 0$ if and only if $\mathbf{Q}_{x,x'} > 0$, the KL divergence is defined as

$$D(\mathbf{P}||\mathbf{Q}) \triangleq \sum_{x\in\mathcal{S}} \gamma(x) \sum_{x'\in\mathcal{S}} \mathbf{P}_{x,x'} \log \frac{\mathbf{P}_{x,x'}}{\mathbf{Q}_{x,x'}}, \tag{5.5}$$

where $\{\gamma(x)\}$ is the stationary distribution of $\mathbf{P}$. Note that $D(\mathbf{P}||\mathbf{Q})$ is the expected KL divergence between the distributions of the next state under $\mathbf{P}$ and $\mathbf{Q}$ conditioned on the current state (i.e., individual rows of $\mathbf{P}$ and $\mathbf{Q}$), where the expectation is with respect to the stationary distribution of $\mathbf{P}$ (i.e., the current state is drawn randomly according to the stationary distribution of $\mathbf{P}$). The proof of the above lower bound follows a similar line of arguments as in Lai and Robbins, 1985 [126].

Asymptotic and order optimal policies: The Lai–Robbins policy was extended by Anantharam, Varaiya, and Walrand, 1987 [15], and shown to preserve its asymptotic optimality. Specifically, a point estimate of the mean $\hat{\mu}_i(t)$ and UCB $\{U_i(t)\}_{t\geq 1}$ can be similarly constructed using large deviation results on the empirical distribution and the empirical transition-count matrix of finite-state Markov chains and under an additional assumption that $\log \mathbf{P}_{x,x'}(\theta)$ is a concave function of θ for all $x, x' \in \mathcal{S}$.

The UCB-α policy given in Algorithm 4.3 can also be extended to handle rested reward processes and maintain its order optimality as shown by Tekin and Liu, 2012 [189]. The only necessary change is to set the parameter α in the UCB-α index to a value sufficiently large as determined by the second largest eigenvalues of the transition matrices as well as the stationary distributions. Prior knowledge on nontrivial bounds on such quantities suffice for setting this exploration constant. The DSEE and ϵ-greedy policies can also be extended without much difficulty.

5.1.2 RESTLESS MARKOV REWARD PROCESSES

Similar to the Bayesian restless bandit discussed in Section 3.3, the frequentist restless bandit model involves reward processes that continue to evolve even when the associated arms are passive. The difference here is that the Markov models governing the state transitions are unknown under both active and passive modes of the arms. To see it another way, the Bayesian restless bandit model discussed in Section 3.3 defines the oracle strategy for the benchmark performance that learning algorithms for the frequentist restless bandit aim to approach.

Analytical characterizations of the optimal policy for the Bayesian restless bandit are intractable. It has been shown to be P-SPACE hard by Papadimitriou and Tsitsiklis, 1999 [162]. Whittle index policy is a good candidate for a proxy. Unfortunately, analytical characterizations of the Whittle index policy are also intractable in general (in Chapter 6 we show a special case where a logarithmic regret order with respect to the Whittle index policy can be achieved).

An alternative is to consider the single-arm policy that always plays the arm with the greatest limiting reward rate, i.e., the proxy oracle for the rested case. This proxy, however, can no longer preserve the same order of performance as the oracle. In fact, the performance gap to the oracle can grow linearly with T. The resulting regret is thus a much weaker measure. This notion of *weak regret* was also adopted by Auer et al., 2003 [24], for bandits with non-stochastic reward processes.

Under the notion of weak regret, the UCB and the DSEE policies have been extended to offer a logarithmic regret order for restless reward processes. The extensions, however, are much more involved than in the rested case. Compared to the i.i.d. and the rested Markovian reward models, the restless nature of arm state evolution requires that each arm be played consecutively for a period of time in order to learn its Markovian reward statistics. The length of each segment of consecutive plays needs to be carefully controlled to avoid spending too much time on a suboptimal arm. At the same time, we experience a transient period each time we switch out and then back to an arm, which leads to potential reward loss with respect to the steady-state behavior of this arm. Thus, the frequency of arm switching needs to be carefully bounded.

These factors are balanced through an epoch structure in the extension of the DSEE policy by Liu, Liu, and Zhao, 2013 [132]. Specifically, the time horizon is partitioned into interleaving exploration and exploitation epochs with carefully controlled epoch lengths. During an exploration epoch, the player partitions the epoch into N contiguous segments, one for playing each of the N arms to learn their reward statistics. During an exploitation epoch, the player plays the arm with the largest sample mean (i.e., average reward per play) calculated from the observations obtained so far. The lengths of both the exploration and the exploitation epochs grow geometrically. The number of arm switchings are thus at the logarithmic order with time. The tradeoff between exploration and exploitation is balanced by choosing the cardinality of the sequence of exploration epochs.

The UCB policy was extended by Tekin and Liu, 2012 [189], through a clever use of the i.i.d. structure of the regenerative cycles of a Markov chain. The basic idea is to play an arm consecutively for a random number of times determined by a regenerative cycle of a particular state and arms are selected based on the UCB index calculated from observations obtained only inside the regenerative cycles (observations obtained outside the regenerative cycles are not used in learning). The i.i.d. nature of the regenerative cycles reduces the problem to the canonical bandit model with i.i.d. reward processes. A different extension of the UCB policy was developed by Liu, Liu, and Zhao, 2013 [132], by introducing a geometrically growing epoch structure. Compared to the extension based on regenerative cycles, this extension avoids the potentially large number of observations not used for learning before the chosen arm enters a regenerative cycle defined by a particular pilot state.

5.1.3 NONSTATIONARY REWARD PROCESSES

When the reward process of each arm is nonstationary with time-varying distributions, the online learning problem adds another dimension. The problem now becomes a moving-target problem. In addition to the tradeoff between exploitation and exploration, the learning algorithm also faces the dilemma of to remember or to forget past observations. On the one hand, learning efficiency hinges on fully utilizing available data. On the other hand, to track the best arm that changes over time, the algorithm needs to be sufficiently agile by forgetting outdated observations that might have become misleading.

Analytical treatments of the problem depends on the model for the temporal variations of the reward processes. Three different models have been considered in the literature: the abrupt-change model, the continuous-drift model, and the total-variation model as discussed below.

The abrupt-change model: Under this model, the set of reward distributions $\{F_i\}_{i=1}^N$ experiences abrupt changes at unknown time instants. In other words, the reward process of each arm is piecewise stationary with unknown change points (note that this includes scenarios where arms change asynchronously). This model is often parameterized by the total number κ of change points over the horizon of length T.

To handle the abrupt changes of the reward distributions, it is necessary to incorporate into the learning policy a mechanism for filtering out potentially obsolete observations; a direct application of learning algorithms designed under the assumption of time-invariant models results in a near-linear regret order. For instance, a direct application of the UCB policy to bandits experiencing two change points results in a regret order no smaller than $T/\log T$ as shown by Garivier and Moulines, 2011 [85].

There are two general approaches to the filtering mechanism: an open-loop approach that is agnostic to specific change points and adapts only to the total number $\boldsymbol{\gamma}_T$ of change points, and a fully adaptive approach that actively detects the change points based on observed rewards and adjusts the arm selections based on the outcomes of the change detection.

Under the open-loop approach, the filtering component employs standard techniques in model tracking: *discounting, windowing, and restarting*. In particular, the discounting UCB (D-UCB) algorithm uses a discounted empirical average and a corresponding confidence bound based on discounted time for computing the arm indexes. The sliding-window UCB (SW-UCB) computes the arm indexes using only most recent observations within a window of a specific length. The discount factor in D-UCB and the window length in SW-UCB are predetermined based on the total number κ of change points and the horizon length T. Both algorithms degenerate to the original UCB policy when $\kappa = 0$, for which the discount factor in D-UCB becomes 1 and the window length in SW-UCB increases to T. The regret analysis adopts a minimax formulation of the change points, focusing on the worst-case in terms of the specific change points for a given κ. It was shown that D-UCB offers a regret order of $O(\sqrt{\kappa T}\log T)$ and SW-UCB a regret order of $O(\sqrt{\kappa T \log T})$. Both are near order-optimal (up to a $\log T$ and $\sqrt{\log T}$ factor, respectively) when the number κ of change points is bounded. This is established through a $O(\sqrt{T})$ lower bound on the regret order for the case of $\kappa = 2$ by Garivier and Moulines, 2011 [85]. When κ grows linearly with T (i.e., the change points have a positive density), a linear regret order is inevitable, since each abrupt change incurs performance loss. Besides these two extreme points in terms of κ, a full characterization of the optimal regret scaling with respect to κ is lacking. The restarting technique employed by Auer et al., 2003 [24] for non-stochastic bandits can be employed here as well, where a standard learning policy can be augmented with a periodic restarting with the restarting period chosen *a priori* based on κ.

The adpative approach employs a change detector to trigger restarts of the learning algorithm. Representative results include Hartland et al., 2007 [100] and Liu, Lee, and Shroff, 2018 [133], that integrate classical change detection algorithms such as the Page-Hinckley Test (PHT) and cumulative sum control chart (CUSUM) test with the UCB policies. In particular, it was shown by Liu, Lee, and Shroff, 2018 [133] that when known lower bounds are imposed on the changes in the reward mean as well as the time lapse between two consecutive change points, the CUSUM-UCB policy offers a regret order of $O(\sqrt{\kappa T \log \frac{T}{\kappa}})$ with the knowledge of κ. Allesiardo and Feraud, 2015 [12] proposed a confidence-bound based test for detecting a switching of the best arm (not every change in reward models results in a switching of the best arm) to restart EXP3, a randomized policy originally developed for nonstochastic bandits (see Section 5.1.4). Referred to as EXP3.R (EXP3 with restarting), this policy achieves a regret order of $O(\tilde{\kappa}\sqrt{T \log T})$, where $\tilde{\kappa} \leq \kappa$ is the number of switchings of the best arm, without the knowledge of $\tilde{\kappa}$ or κ. Assuming a specific stochastic model for the change points, Mellor and Shapiro, 2013 [149] integrated a Bayesian change point detector into Thompson Sampling.

The continuous-drift model: The abrupt-change model focuses on temporal variations that are significant in magnitude—with the minimum shift in reward mean bounded away from zero—yet sparse in occurrence (with the number κ of change points growing sublinearly with T). Slivkins and Upfal, 2008 [180] introduced a Brownian bandit model to capture continuous and gradual drift of the reward models as often seen in economic applications (e.g., stock price, supply and demand). In particular, the temporal dynamics of the reward mean of each arm $\mu_i(t)$ follow a Brownian motion with rate σ_i^2. To avoid unbounded drift, the Brownian motions are restricted to an interval with reflecting boundaries. The objective is to characterize the regret *per play* in terms of the volatilities $\{\sigma_i^2\}_{i=1}^N$ of the arm reward models. The constant lower bound on the regret per play established by Slivkins and Upfal, 2008 [180] implies a linear growth rate of the cumulative regret in T.

The total-variation model: Besbes, Gur, and Zeevi, 2014 [36] considered a temporal variation model that imposes an upper bound on the total variation of the expected rewards over the entire horizon. The total variation is defined as

$$\upsilon_T = \sum_{t=1}^{T-1} \max_{i=1,\dots,N} |\mu_i(t) - \mu_i(t+1)|.$$

The objective is to characterize the minimax regret order over the set of reward models with a total variation no greater than a variation budget υ_T. More specifically, the regret of a policy π is given by

$$R_\pi(T;\upsilon_T) = \sup_{\mathcal{F}_{\upsilon_T}} \mathbb{E}\left[\sum_{t=1}^{T} \mu_*(t) - \mu_{\pi(t)}(t)\right],$$

where $\mathcal{F}_{\upsilon_T}$ denotes the set of distribution sequences $\{F_i(t)\}_{i=1,\ldots,N;\, t=1,\ldots,T}$ with bounded support in $[0, 1]$ satisfying the variation budget υ_T and $\mu_*(t)$ is the maximum expected reward at time t. Besbes, Gur, and Zeevi, 2014 [36] established a regret lower bound of $\Omega((N\upsilon_T)^{1/3}T^{2/3})$ and showed that EXP3 with periodic restarting achieves the optimal order up to a logarithmic term in N. The restarting period is predetermined based on the variation budget υ_T. In other words, it is an open-loop policy in terms of its handling of model variations.

The total variation model includes abrupt changes and continuous drifts as special cases, with a proper translation from the number κ of changes and the volatilities $\{\sigma_i^2\}_{i=1}^N$ to the total variation budget υ_T. The characterization of the minimax regret over this broader class, however, does not apply to more specific models. Results with a minimax nature, when developed under a broader model class, often sacrifice sharpness. When more information is available, tighter lower bounds and better policies can be tailor-made for the specific model class.

5.1.4 NONSTOCHASTIC REWARD PROCESSES: ADVERSARIAL BANDITS

Under an adversarial bandit model, the reward process of an arm i is an arbitrary unknown deterministic sequence $\{x_i(t)\}_{t\geq 1}$ with $x_i(t) \in [0, 1]$ for all i and t. In this case, one may question what can actually be learned from past reward observations if the past and potential future rewards are not bound by a common stochastic model and have no structure.

Indeed, if the objective is to approach the performance of an oracle who knows the entire reward process of every arm *a priori* and plays optimally based on this knowledge, the learning task is hopeless, and a thus-defined regret will grow linearly with T.

Weak regret, however, leads to a well-formulated problem. This corresponds to an oracle who can only choose one fixed arm to play over the entire time horizon. Under the objective of minimizing the weak regret, what the player is trying to learn is which arm has the largest *cumulative* reward rather than trying to catch the largest reward at each time instant. Intuitively, under the assumption of bounded reward, the former is possible as past reward observations become increasingly more informative for learning the largest cumulative reward as time goes.

Besides adjusting the benchmark in the regret definition, the adversarial bandit model also necessitates randomized strategies. A deterministic policy can be easily defeated by adversarially chosen reward sequences, leading to a linear order of even the weak regret.

Under the minimax approach, the weak regret of an arm selection policy π is given by

$$R_\pi(T) = \max_{\{x_i(t)\}_{t\geq 1, i=1,\ldots,N}} \left\{ \max_{i=1,\ldots,N} \left(\sum_{t=1}^{T} x_i(t) \right) - \mathbb{E}\left[\sum_{t=1}^{T} x_{\pi_t}(t) \right] \right\}, \tag{5.6}$$

where the expectation is over the internal randomness of policy π. In other words, the performance of a policy π is measured against the worst reward sequences $\{x_i(t)\}_{t\geq 1, i=1,\ldots,N}$.

The notion of the weak regret in the bandit setting corresponds to the *external regret* commonly adopted in game theory, where the utility of a strategy is measured against the best *single* action in hindsight. In the context of games, a popular learning algorithm for minimizing external regret is the *multiplicative weights* method. It maintains a weight $w_a(t)$ for each action a (which corresponds to an arm in the bandit setting) as an indicator of the potential payoff of this action. Specifically, the weight is given by

$$w_a(t) = e^{\eta(t)\sum_{j=1}^{t} x_a(j)}, \tag{5.7}$$

where $x_a(j)$ is the jth reward sample from taking action a and $\{\eta(t)\}_{t\geq 1}$, referred to as the learning rate (in the role of a step-size), is a positive and monotonically decreasing sequence. At time $t+1$, the algorithm chooses a random action a at time t with the following probability:

$$p_a(t+1) = \frac{w_a(t)}{\sum_{a'} w_{a'}(t)}. \tag{5.8}$$

The name of the algorithm comes from the fact that the weight $w_a(t)$ is updated multiplicatively under a constant learning rate η, i.e., $w_a(t) = w_a(t-1)e^{\eta x_a(t)}$.

The multiplicative-weights method was developed under full-information feedback where the rewards of all actions a player could have taken are revealed at the end of each decision period. For minimizing the weak regret in the adversarial bandit setting, Auer et al., 2003 [24] modified this algorithm in order to handle the change in the feedback model from full-information to bandit feedback. Referred to as Exploration-Exploitation with Exponential weights (EXP3), this bandit algorithm differs from the multiplicative-weight algorithm in two aspects: (i) correcting the bias (caused by the incompleteness of the bandit feedback) in the estimated reward of the chosen action by dividing the probability of selecting it; and (ii) mixing the action distribution with a uniform distribution for exploration to ensure that sufficient information is obtained for all arms under the bandit feedback. It was later noted by Stoltz (2005) [183] that this uniform exploration is unnecessary. The implementation given in Algorithm 5.5 is without this uniform exploration. Another modification in Algorithm 5.5 is that the unbiased reward estimate is obtained through an unbiased estimate of the loss $1 - x_i(t)$ to mitigate the issue of large variance of the estimates associated with arms played with a small probability (see a detailed discussion in Lattimore and Szepesvári, 2019 [129]).

The weak regret of EXP3 was shown to be $O(\sqrt{TN \log N})$. An $\Omega(\sqrt{TN})$ lower bound on the weak regret follows directly from the lower bound on the minimax regret in the stochastic setting as given in Theorem 4.7. Specifically, if the regret averaged over reward sample paths generated under a certain set of arm distributions is $\Omega(\sqrt{TN})$, there must exist a sample path for which the (weak) regret has an order that is no less than $\sqrt{TN}$. The $\log N$ gap between the regret order of EXP3 and the lower bound was closed by Audibert and Bubeck, 2009 [19] by considering a class of functions more general than the exponential function for weights.

Other notions of regret defined with respect to a benchmark policy with a given hardness constraint on the number of arm switches were also considered in Auer et al., 2003 [24].

Algorithm 5.5 The EXP3 Algorithm

```
Initialization: set $v_i(0) = 0$ for $i = 1, \ldots, N$.
for $t = 1$ to $T$ do
    Set $\eta(t) = \sqrt{\frac{\log N}{tN}}$.
    Calculate the distribution for arm selection:
        $p_i(t) = \frac{\exp\left(\eta(t) v_i(t-1)\right)}{\sum_{j=1}^{N} \exp\left(\eta(t) v_j(t-1)\right)}$.
    Pull arm $i_t$ chosen randomly according to the probability distribution $[p_1(t), \ldots, p_N(t)]$.
    Receive reward $x_{i_t}(t)$ from arm $i_t$.
    Calculate $v_i(t)$ for all $i$:
        $v_i(t) = v_i(t-1) + \left(1 - \frac{1}{p_i(t)}(1 - x_{i_t}(t))\mathbb{I}[i = i_t]\right)$.
end for
```

5.2 VARIATIONS IN THE ACTION SPACE

5.2.1 LARGE-SCALE BANDITS WITH STRUCTURED ACTION SPACE

The canonical bandit model assumes independence across arms. In this case, reward observations from one arm do not provide any information on the quality of other arms. A linear growth rate with N of the problem-specific regret is expected given that every arm needs to be sufficiently explored. The main focus of the bandit theory has thus been on the regret growth rate with T, which measures the learning efficiency over time.

Many emerging applications lead to bandit problems with a massive and sometimes an uncountable number of arms. Consider, for example, adaptive routing in a network with unknown and stochastically varying edge weights (see Section 6.1.2). Each path from the source node to the destination node constitutes an arm, and the number of possible paths can grow exponentially with the network size. In the application of ads display on search engines, the collection of all ads is the set of arms with reward measured by the number of clicks. In political campaign or targeted marketing of a new product, arms are individuals in a massive population. When casting stochastic optimization of a random unknown function as a bandit problem, the arm set is given by the domain of the objective function, consisting of uncountable alternatives in a potentially high-dimensional Euclidean space. Developed under the assumption of independent arms and relying on exploring every arm sufficiently often, learning policies for the canonical bandit model no longer offer appealing or even feasible solutions, especially in the regime of $N > T$.

The key in handling a large number of arms is to fully exploit certain known structures and relations among arms that naturally arise in specific applications. For instance, in network optimization problems such as adaptive routing, the large number of arms (i.e., the paths) are dependent through a much smaller number of unknowns (i.e., the edge weights). In many social-economic applications, arms have natural similarity and dissimilarity relations that bound the difference in their expected rewards. For instance, in recommendation systems and information retrieval, products, ads, and documents in the same category (more generally, close in a certain feature space) have similar expected rewards. At the same time, it may also be known *a priori* that some arms have considerably different mean rewards, e.g., news with drastically different opinions, products with opposite usage, documents associated with key words belonging to opposite categories in the taxonomy. Such known arm relations offer the possibility of a sublinear regret growth rate with N, indicating that the best arm can be learned by trying out only a diminishing fraction of arms when N approaches to infinity.

A graphical random field representation of action space: The structure of the action space of a bandit model can be represented by a *graphical random field*, which is a generalization of graphical models. A random field is a collection of random variables indexed by elements in a topological space. It generalizes the concept of stochastic process in which the random variables are indexed by real or discrete values referred to as time. A graphical random field $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set $\mathcal{V}$ of vertices representing random variables and a set $\mathcal{E}$ of potentially directed edges representing the presence of a certain relation between connected random variables. The vertex set $\mathcal{V}$ can be partitioned into two subsets, one consists of N random variables $\{X_i\}_{i=1}^{N}$ representing the random reward of each of the N arms,[1] the other consists of M latent random variables $\{Y_j\}_{j=1}^{M}$ that determine or influence the first set of random variables. The distributions of both sets of random variables are unknown. A realization of X_i (i.e., a reward) is observed once this arm is engaged. The latent variables are in general (but not always) unobservable. The edges represent certain relations across the random variables associated with the vertices. The relations represented by edges can be categorized into two classes—realization-based relations and ensemble-based relations—as detailed blow.

Realization-based arm relations: Realization-based relations capture how the value (i.e., a realization) of a random variable is affected by the values of its neighbors in the graphical random field. Most existing reward structures considered in the bandit literature are special cases of this class. Representative models include the combinatorial bandits considered by Gai, Krishnamachari, and Jain, 2011 [82], Liu and Zhao, 2012 [137], Chen, Wang, and Yuan, 2013 [60], and Kveton et al., 2015 [124], linearly parameterized bandits by Dani, Hayes, and Kakade,

[1] This model also applies to bandits with a continuum set of arms due to the general definition of random fields.

2008 [72], Rusmevichientong and Tsitsiklis, 2010 [173], Abbasi-Yadkori, Pal, and Szepesvari, 2011 [1], and spectral bandits for smooth graph functions by Valko et al., 2014 [199] and Hanawal and Saligrama, 2015 [97]. Specifically, the resulting graph is a bipartite graph with the latent variables $\{Y_j\}_{j=1}^M$ and the arm variables $\{X_i\}_{i=1}^N$ as the partite sets (see Figure 5.1). The value of X_i is a known deterministic function of its latent neighbors:

$$X_i = h_i(Y_j : (Y_j, X_i) \in \mathcal{E}). \tag{5.9}$$

In particular, for combinatorial bandits, the deterministic functions h_i are often the sum[2] of the input:

$$X_i = h_i(Y_j : (Y_j, X_i) \in E) = \sum_{j:(Y_j, X_i)\in\mathcal{E}} Y_j. \tag{5.10}$$

Consider the specific application of adaptive routing which leads to a combinatorial bandit. The latent variables $\{Y_j\}_{j=1}^M$ are the random weights of the M edges in the network, and the arm variables $\{X_i\}_{i=1}^N$ are the path weights from the source to the destination. The path weight (i.e., the reward/cost for playing an arm) is the sum of the edge weights along this path. In the work by Liu and Zhao, 2012 [137], the latent variables (i.e., the individual edge weights) are not observable. In the work by Gai, Krishnamachari, and Jain, 2011 [82], Chen, Wang, and Yuan, 2013 [60], and Kveton et al., 2015 [124], it is assumed that all latent variables controlling the chosen arm are also observed. The latter setting is referred to as the combinatorial semi-bandit to differentiate it from the bandit setting where only the reward of the chosen arm, not the constituent latent variables, is observed.

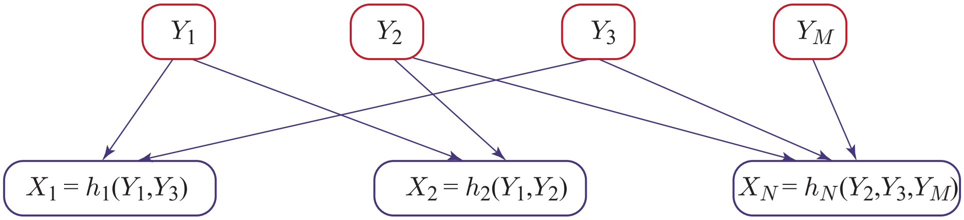


Figure 5.1: A graph representation of a structured action space.

For linear bandits considered by Dani, Hayes, and Kakade, 2008 [72], Rusmevichientong and Tsitsiklis, 2010 [173], and Abbasi-Yadkori, Pal, and Szepesvari, 2011 [1], the bipartite graph is complete, i.e., there is an edge between every pair of (Y_j, X_i). Each deterministic function h_i is specified by an M-dimensional real vector $\mathbf{w}_i$ that is known, and the reward of this arm/action is given by a weighted sum of all latent variables[3]: $X_i = \sum_{j=1}^M \mathbf{w}_i(j) Y_j$. The spectral

[2] In the work by Chen, Wang, and Yuan, 2013 [60], nonlinear functions are allowed provided that the nonlinear function is commutative with the expectation operation.

[3] Some linear bandit models considered in the literature also allow a zero-mean random noise term added to each arm variable. This, however, does not change the reward structure given that only the mean values of the arms matter under the objective of expected total reward.

bandit considered by Valko et al., 2014 [199] and Hanawal and Saligrama, 2015 [97], is a special case of the linear bandits with $M = N$ and the weight vectors $\mathbf{w}_i$ given by the ith eigenvector of a known fixed graph.

More general realization-based relations (for example, h_i can be a random function or specifies the conditional distribution) can be modeled under the graphical random field representation for studying large-scale bandits with a structured action space and how various aspects of the structure affect the learning efficiency both temporally (in T) and spatially (in N).

Ensemble-based arm relations: Ensemble-based relations capture arm similarities in their ensemble statistics—in particular, the mean—rather than probabilistic dependencies in their realizations.

A first such example is the continuum-armed bandit problem first formulated by Agrawal, 1995b [5] under the minimax approach.

Definition 5.1 *The Continuum-Armed Bandit Model:*
Let $\mathcal{S} \subset \mathcal{R}$ be a bounded set, representing a continuum set of arms. Playing arm $s \in \mathcal{S}$ generates a random reward drawn from an unknown distribution f_s, i.i.d. over time. Let $\mu(s)$ denote the expected reward of arm s. It is assumed that $\mu(s) : \mathcal{S} \to \mathcal{R}$ is uniformly locally Lipschitz with constant L $(0 \leq L < \infty)$, exponent α $(0 < \alpha \leq 1)$, and restriction δ $(\delta > 0)$, i.e., for all $s, s' \in \mathcal{S}$ satisfying $|s - s'| \leq \delta$, we have

$$|\mu(s) - \mu(s')| \leq L|s - s'|^{\alpha}. \tag{5.11}$$

Let $\mathcal{L}(\alpha, L, \delta)$ denote the class of all such functions. The objective is an arm selection policy that minimizes the regret in the minimax sense over $\mathcal{L}(\alpha, L, \delta)$. Specifically, the worst-case regret of policy π is defined as

$$R_\pi(T) \triangleq \sup_{\mu(s) \in \mathcal{L}(\alpha, L, \delta)} \sum_{t=1}^{T} (\mu_* - \mu(s_\pi(t))),$$

where $\mu_* = \sup_{s \in \mathcal{S}} \mu(s)$ and $s_\pi(t)$ is the arm chosen by policy π at time t.

The smoothness of the mean reward function $\mu(s)$ as imposed by the uniformly locally Lipschitz condition is the key to handling the continuum set of arms, making it possible to approach the optimal arm by a sparse sampling the continuum set. This bandit model is also known as the Lipschitz bandit.

An algorithm (UniformMesh) was developed by Kleinberg, 2004 [115], based on an open-loop approximation of the mean reward function $\mu(s)$. Intuitively, the smoothness of the mean reward makes it possible to approximate the continuum-armed bandit problem with a finite-armed bandit problem by a discretization of the arm set. It was shown by Kleinberg, 2004 [115] that this strategy based on an open-loop uniform discretization of the arm set offers a regret order of $O(T^{2/3} \log^{1/3}(T))$, matching the lower bound of $\Omega(T^{2/3})$ up to a sublogarithmic factor.

It might appear quite surprising that such a simple open-loop reduction of the continuum-armed bandit to a finite-armed bandit suffices for near-optimal regret performance. The reason behind this is the minimax criterion that focuses on the performance under the worst possible mean-reward functions in $\mathcal{L}(\alpha, L, \delta)$. Such worst possible payoff functions are those that are the least smooth (i.e., achieving the upper bound in the Lipschitz condition (5.11)) around the maxima. Thus, a uniform discretization of the arm set with a predetermined level of refinement (given by the Lipschitz exponent α) suffices.

Intuition suggests that the hardness of a continuum-armed bandit is determined by the smoothness of the payoff function around the maxima, not the global parameters that bound the smoothness over the entire domain. It is thus desirable to have learning strategies that can adapt automatically to the given payoff function and take advantage of its higher-order smoothness around the maxima if present to achieve better problem-specific performance. The same intuition also suggests that the global Lipschitz condition can be relaxed to local smoothness properties around the maxima of the payoff function.

These objectives call for adaptive strategies that successively refine arm discretization based on past observations. The resulting discretization of the arm set is nonuniform, with more observation points around the maxima and the refinement adapting automatically to the local smoothness around the maxima. Representative studies along this line include Auer, Ortner, and Szepesvari, 2007 [25], Bubeck, Munos, Stoltz, and Szepesvari, 2011 [48], Cope, 2009 [69], and Kleinberg, Slivkins, and Upfal, 2015 [117], assuming various assumptions on the local smoothness of the payoff function around the maxima, leading to different regret orders.

The continuum-armed bandit over a compact set of the real line has been extended to high-dimension metric space by Kleinberg, Slivkins, and Upfal, 2015 [117], and generic measurable space by Bubeck, Munos, Stoltz, and Szepesvari, 2011 [48]. Another direction is to consider more restrictive payoff functions, including linear and convex functions for the application of online stochastic optimization (see Dani, Hayes, and Kakade, 2008 [72], Agarwal et al., 2013 [2], and references therein). The issue of unknown parameters in the smoothness properties of the payoff function has been addressed by Bubeck, Stoltz, and Yu, 2011 [49], Slivkins, 2011 [177], Minsker, 2013 [150], Valko, Carpentier, and Munos, 2013 [198], and Bull, 2015 [53].

Other bandit models exploiting ensemble-based arm structures include the taxonomy bandit (Slivkins, 2011 [177]), the unimodal bandit (Combes and Proutiere, 2014 [66]), and the unit-interval-graph (UIG) bandit (Xu et al., 2019 [214]). In the taxonomy bandit model, statistical similarities across arms are captured by a tree structure that encodes the following ensemble relation: arms in the same subtree are close in their mean rewards. In unimodal bandits, the ensemble-based arm relation is represented by a graph with the following property: from every sub-optimal arm, there exists a path to the optimal arm along which the mean rewards increase. The UIG bandit model exploits not only statistical similarities, but also statistical dissimilarities across arms, with the similarity-dissimilarity relations represented by a unit interval graph that may be only partially revealed to the player. A general formulation of structured bandits was proposed in Combes, Magureanu, and Proutiere, 2017 [68], which includes a number of bandit models (e.g., linear bandits, Lipschitz bandits, unimodal bandits, and UIG bandits) as special cases. The learning policy developed there, however, was given only implicitly in the form of a linear program that needs to be solved at every time step. For certain bandit models (e.g., the UIG bandit), the linear program does not admit polynomial-time solutions (unless P=NP). We see again here the tension between the generality of the problem model and the efficiency of the resulting solutions.

5.2.2 CONSTRAINED ACTION SPACE

There are several studies on bandit models where not all arms are available for activation at all times. This can be determined by either a given exogenous process or through self control to satisfy a certain budget constraint.

An example of the first type is the so-called sleeping bandit studied by Kleinberg, Miculescu-Mizil, and Sharma, 2008 [116]. Specifically, the set of available arms is time-varying, resulting in time varying identities of the best arm depending on the set of available arms. This variant can be quite easily dealt with. It is shown that a simple modification of the UCB policy that chooses the arm with the greatest index among the currently available ones achieves the optimal regret order. This problem was also studied under the non-stochastic setting with adversarial reward processes by Kleinberg, Miculescu-Mizil, and Sharma, 2008 [116] and Kanade, McMahan, and Bryan, 2009 [107].

The second type is the so-called Knapsack bandit or bandit with a budget (see Tran-Thanh et al., 2010 [191]; Tran-Thanh et al., 2012 [192]; Badanidiyuru, Kleinberg, and Slivkins, 2013 [28]; Jiang and Srikant, 2003 [103]; Combes, Jiang, and Srikant, 2015 [67]). Specifically, the horizon length T (which can be viewed as a budget constraint) is replaced with a more general form of budget B for arm activation. Each activation of an arm, in addition to generating random rewards, consumes a certain amount of budget that can be arm-dependent. The objective is to minimize the cumulative regret over the course till the depletion of the total budget. The focus is similarly on order optimality as the budget B tends to infinity. The learning algorithms often employ UCB-type indices with the quality of an arm measured by the reward-to-cost ratio.

5.3 VARIATIONS IN THE OBSERVATION MODEL

5.3.1 FULL-INFORMATION FEEDBACK: THE EXPERT SETTING

The bandit feedback reveals only the reward of the chosen arm. It is this absence of observations from unselected arms that dictates the need for exploration. In certain applications, however, the rewards of unselected arms, although not realized, can be observed. Consider a stylized portfolio selection problem that decides periodically (e.g., every month) which financial advisor to follow. Although only the gain from the selected portfolio is accrued, the performance of the portfolios suggested by other advisors is also observed and can be used in choosing the next action. The objective is to minimize the cumulative loss (i.e., regret) against the performance of the best advisor in the given set. This sequential learning problem with full-information feedback is often referred to as prediction with expert advice, or the expert setting in short.

Since all arms are observed at all times, exploration is no longer a consideration when deciding which arm to pull. Consequently, under i.i.d. reward processes, the current action has no effect on the future. The optimal action at each time is to exploit fully for maximum instantaneous gain. Since the best arm can be identified with probability 1 within finite time, a bounded regret can be achieved, and a simple follow-the-leader strategy suffices.

While the stochastic version of the expert setting admits simple solutions, its nonstochastic counterpart where the reward sequence of each expert/arm is deterministic and adversarially designed presents sufficient complexity. It has been studied extensively within a game-theoretic framework since the 1950s with the pioneering work of Blackwell, 1956 [40] and Hannan, 1957 [98]. A comprehensive coverage of this subject can be found in the book by Cesa-Bianchi and Lugosi, 2006 [58].

5.3.2 GRAPH-STRUCTURED FEEDBACK: BANDITS WITH SIDE OBSERVATIONS

The bandit feedback and the full-information feedback are at the two ends of the spectrum in terms of the information available for learning after each chosen action. A general observation model that includes these two as special cases is as follows. At each time t, after pulling arm i, the rewards of arms in a set $\Psi(t, i)$ are observed. This set may depends on not only the action i, but also the time index t. The bandit feedback and the expert feedback correspond to, respectively, $\Psi(t, i) = \{i\}$ and $\Psi(t, i) = \{1, 2, \ldots, N\}$ for all t and i.

This observation model can be represented by a directed graph $\mathcal{G}_t$ for each t. Specifically, each vertex of $\mathcal{G}_t$ represents an arm. A directed edge from vertex i to vertex j exists if and only if pulling arm i at time t leads to an observation of arm j, i.e., if and only if $j \in \Psi(t, i)$. It is easy to see that under the bandit feedback, $\mathcal{G}_t$ consists of only self-loops and is time invariant. The expert setting corresponds to a time-invariant graph with all N^2 directed edges.

Referred to as bandits with side observations, this online learning problem with graph-structured feedback was first introduced by Mannor and Shamir, 2011 [144], within a nonstochastic framework. The stochastic version of the problem was studied by Caron et al., 2012 [55] and Buccapatnam, Eryilmaz, and Shroff, 2014 [52].

The central issue under study is how the graph-structured feedback affect the scaling of regret with respect to the number N of arms. It has been shown that under the nonstochastic formulation, it is the independence number of the observation graph that determines the regret scaling with N. Under the stochastic formulation, regret can be made to scale with the clique partition number or the size of the minimum dominating set of the observation graph. The issues of learnability when certain self-loops are missing and when the observation graph is time varying and never fully revealed to the player are studied by Alon et al., 2015 [13] and Cohen, Hazan, and Koren, 2016 [65].

5.3.3 CONSTRAINED AND CONTROLLED FEEDBACK: LABEL-EFFICIENT BANDITS

In certain applications, obtaining feedback can be costly. This can be captured by imposing a constraint on the total number of observations that can be obtained over the entire horizon of length T. Specifically, at each time t, after choosing an arm to pull, the player decides whether to query the reward obtained from the current pull of the chosen arm, knowing that the total number of such queries cannot exceed a given value. The resulting formulation is referred to as label-efficient bandits, following the terminology in learning from labeled data. Such constrained and controlled feedback can also be incorporated into the expert setting, with each query producing reward observations of all arms. A fully controlled feedback model would be to allow the player to choose, at each time, which subset of arms to observe, while complying with a constraint on the total number of labels summed over time and across all arms. In other words, the player designs the sequence of feedback graphs $\{\mathcal{G}_t\}_{t\geq 1}$ that provides the most information under the given constraint. An interesting question is whether this feedback model may result in a decoupling of exploration and exploitation, with the feedback graphs designed purely for exploration and the arm pulling actions chosen for pure exploitation. It appears that this model has not been studied in the literature.

Existing studies on constrained and controlled feedback (both the bandit and the expert settings) all adopt the nonstochastic/adversarial formulation (see, for example, Helmbold and Panizza, 1997 [101]; Cesa-Bianchi, Lugosi, and Stoltz, 2005 [56]; Allenberg et al., 2006 [11]; Audibert and Bubeck, 2010 [20]). The deterministic nature of the reward sequences and the focus on the worst-case performance (i.e., minimax regret) render sophisticated query strategies unnecessary. A simple open-loop control of the queries suffices to achieve the optimal minimax regret order: simply toss a coin to determine whether to query at each time with the bias of the coin predetermined by the feedback constraint and the horizon length. Whether the stochastic setting under the uniform dominance objective would demand a more sophisticated query strategy appears to be an unexplored direction.

5.3.4 COMPARATIVE FEEDBACK: DUELING BANDITS

The traditionally adopted observation model and its variants discussed above all assume that an absolute quantitative measure of arm payoffs exists and can be observed on demand. In certain applications, however, this assumption may not hold. For instance, in determining the quality/relevance of an image or a rank-listed documents for a given query, it is difficult to obtain unbiased response from users in terms of a quantitative value. However, the preference of the user when comparing two options can be reliably obtained. This motivates the formulation of dueling bandits, in which observations are binary preference outcomes for comparing a pair of arms.

One way to formulate a dueling bandit problem is to preserve the reward model of the canonical bandit problem and change only the observation model. Specifically, a quantitative measure for the rewards offered by each arm exists, although it cannot be observed. At each time t, the player chooses a pair of arms i and j ($i = j$ is allowed) and obtains a binary feedback on whether arm i is better than arm j, which is determined by whether a random reward X_i from arm i drawn from an unknown distribution F_i is greater than that of X_j from arm j. The binary feedback is thus given by a Bernoulli random variable with parameter $\mathbb{P}_{F_i \times F_j}[X_i > X_j]$. Depending on the applications, the actual reward accrued by the player from choosing a pair of arms can be defined as the maximum, the minimum, or the average of the random rewards offered by the chosen pair. Regret can then be defined in the same way as in the canonical bandit model.

Another approach is to adopt directly a probabilistic model for pairwise comparisons of arms without binding it to or assuming the existence of a generative reward model characterized by $\{F_i\}_{i=1}^N$. The underlying probabilistic model for the comparative feedback is given by an $N \times N$ preference matrix $\mathbf{P}$ whose (i, j)th entry $p_{i,j}$ indicates the probability that arm i is preferred over arm j. It is natural to set $p_{i,i} = \frac{1}{2}$ and $p_{i,j} = 1 - p_{j,i}$. We say arm i beats arm j when $p_{i,j} > \frac{1}{2}$.

Under this formulation, it is not immediately obvious how to define regret. The first issue is how to define the best arm when a linear ordering may not exist (e.g., consider sports teams and the following scenario: A beats B, B beats C, and C beats A). The second issue is how to define the performance loss of the chosen pair of arms with respect to the best arm when a quantitative measure of arm payoffs does not exist.

Defining the winner based on (pairwise) preferences is a rich subject with a long history in social science, political science, and psychology (see, for example, Arrow, 1951 [17]). Two notable definitions are the Condorset winner (one that beats all others, which may not always exist) and the Copeland winner (one that beats the most number of candidates). Once the best arm i_* is defined, the performance loss of a chosen pair of arms (i, j) with respect to the best arm i_* is often defined as various functions of $(p_{i_*,i},\ p_{i_*,j})$ or the normalized Copeland scores (i.e., the fraction of candidates inferior to the one in question) of i_*, i, and j. The choice of the loss measure bears less significance on the algorithm design and regret analysis since they all have similar dependency (with bounded difference) on the number of times a suboptimal arm is chosen.

Existing studies on dueling bandits all take the second formulation that directly assumes a preference matrix $\mathbf{P}$. Various assumptions on $\mathbf{P}$ (e.g., strong stochastic transitivity, stochastic triangle inequality) and different definitions of the best arm and loss measures have been considered. We refer the reader to a few representative studies, see Yue et al., 2012 [216], Zoghi et al., 2014 [220], Komiyama et al., 2015 [119], Zoghi et al., 2015 [221], and Wu and Liu, 2016 [213].

The first formulation that builds upon a generative reward model becomes a special case of the second formulation when the preference probability $p_{i,j}$ is given by $p_{i,j} = \mathbb{P}_{F_i \times F_j}[X_i > X_j]$. An unexplored question is whether the generative reward model behind the comparative feedback offers additional structure for improved learning efficiency.

5.4 VARIATIONS IN THE PERFORMANCE MEASURE

5.4.1 RISK-AVERSE BANDITS

The canonical bandit model targets at maximizing the *expected* return. In many applications such as clinical trials and financial investment, the risk and uncertainty associated with the chosen actions need to be balanced with the objective of high returns in expectation. This calls for risk-averse learning.

Risk measures: The notions of risk and uncertainty have been widely studied, especially in economics and mathematical finance. There is no consensus on risk measures, and likely there is no one-size-fits-all. Different applications demand different measures, depending on which type of uncertainty is deemed as risk.

A widely adopted risk measure is *mean-variance* introduced by Nobel laureate economist Harry Markowits in 1952 [147]. Specifically, the mean-variance $\mathrm{MV}(X)$ of a random variable X is given by

$$\mathrm{MV}(X) = \mathrm{Var}(X) - \rho\, \mu(X), \tag{5.12}$$

where $\mathrm{Var}(X)$ and $\mu(X)$ are, respectively, the variance and the mean of X, the coefficient $\rho > 0$ is the risk tolerance factor that balances the two objectives of high return and low risk. The definition of mean-variance can be interpreted as the Lagrangian relaxation of the constrained optimization problem of minimizing the risk (measured by the variance) for a given expected return or maximizing the expected return for a given level of risk. Its quadratic scaling captures the natural inclination of human decision makers that favor less risky options when the stakes are high.

Mean-variance is defined with respect to a random variable. To adopt this risk measure in bandit models, the first question is how to define the mean-variance of a *random sequence* $\{X_{\pi(t)}(t)\}_{t=1}^{T}$ of rewards obtained under policy π. Three definitions exist in the literature, which we refer to as the *Global Risk Constraint (GRC)*, *Local Risk Constraint (LRC)*, and *Empirical Risk Constraint (ERC)*.

Under GRC, risk constraint is imposed on the total reward seen at the end of the time horizon. The mean-variance of the random reward sequence is thus defined as the mean-variance of the sum of the rewards:

$$\begin{aligned} \mathrm{MV}_G\left(\left\{X_{\pi(t)}(t)\right\}_{t=1}^T\right) &= \mathrm{MV}\left(\sum_{t=1}^T X_{\pi(t)}(t)\right) && (5.13)\\ &= \mathrm{Var}\left(\sum_{t=1}^T X_{\pi(t)}(t)\right) - \rho\,\mathbb{E}\left(\sum_{t=1}^T X_{\pi(t)}(t)\right). && (5.14) \end{aligned}$$

This risk measure is more suitable in applications such as retirement investment where the decision maker is more concerned with the final return but less sensitive to the fluctuations in the intermediate returns.

Under LRC, risk constraint is imposed on the random reward obtained at each time. The mean-variance of the random reward sequence is defined as the sum of the mean-variances over time:

$$\begin{aligned} \mathrm{MV}_L\left(\left\{X_{\pi(t)}(t)\right\}_{t=1}^T\right) &= \sum_{t=1}^T \mathrm{MV}\left(X_{\pi(t)}(t)\right) && (5.15)\\ &= \sum_{t=1}^T \mathrm{Var}\left(X_{\pi(t)}(t)\right) - \rho \sum_{t=1}^T \mathbb{E}\left[X_{\pi(t)}(t)\right]. && (5.16) \end{aligned}$$

This risk measure is more suitable in applications such as clinical trial where the risk in each chosen action needs to be constrained.

Under ERC, risk manifests in the inter-temporal fluctuation in the realized reward sample path and is measured by the *empirical* variance. For a given reward sample path $\{x(t)\}_{t=1}^T$, its empirical mean-variance is defined as

$$\begin{aligned} \mathrm{MV}_E\left(\{x(t)\}_{t=1}^T\right) &= \overline{\mathrm{Var}}\left(\{x(t)\}_{t=1}^T\right) - \rho\,\overline{\mu}\left(\{x(t)\}_{t=1}^T\right) && (5.17)\\ &= \sum_{t=1}^T \left(x(t) - \frac{1}{T}\sum_{t=1}^T x(t)\right)^2 - \rho \sum_{t=1}^T x(t), && (5.18) \end{aligned}$$

where the first term is the empirical variance and the second term the empirical mean, both without the normalization term of $1/T$. Not normalizing with the horizon length allows us to preserve the same relation of the performance measure with the horizon length T as seen in the canonical bandit models as well as in the risk-averse bandits under GRC and LRC. We can then compare the regret scaling behaviors in T on an equal footing.

Averaging over all sample paths gives the empirical mean-variance of the random reward sequence:

$$\mathrm{MV}_E\left(\left\{X_{\pi(t)}(t)\right\}_{t=1}^T\right) = \mathbb{E}_\pi\left[\mathrm{MV}\left(\{x(t)\}_{t=1}^T\right)\right], \qquad (5.19)$$

where $\mathbb{E}_\pi$, as usual, denotes the expectation over the stochastic processes $\{X_{\pi(t)}(t)\}_{t=1}^T$ induced by the policy π. This risk measure, first introduced by Sani, Lazaric, and Munos, 2012 [175], directly targets at inter-temporal fluctuations in each realized return process. Such inter-temporal variations are commonly referred to as volatility in portfolio selection and investment (see French, Shwert, and Stambaugh, 1987 [80]) or risk for financial security (Bradfield, 2007 [42]).

Regret decomposition: The above three mean-variance measures are much more complex objective functions than the expected total reward measure in the risk-neutral canonical bandit model. The first challenge we face is that finding the optimal policy under a known model is no longer straightforward or even tractable.

Consider first the ERC model. It can be shown that playing the arm i_* that has the smallest mean-variance may not be optimal (see a counter example in Vakili and Zhao, 2016 [196]). To see this, the key is to notice that the variance term (i.e., the first term on the right-hand side of (5.18)) in the empirical mean-variance measure is with respect to the sample mean calculated from rewards obtained from all arms. When the remaining time horizon is short and the current sample mean is sufficiently close to the mean value of a suboptimal arm $j \neq i_*$, it may be more rewarding (in terms of minimizing the mean-variance) to play arm j rather than arm i_*. In general, the oracle policy π_* has complex dependencies on the reward distributions $\mathbf{F}$ and the horizon length T, and a general analytical characterization appears to be intractable.

This difficulty can be circumvented again by finding a proxy for the oracle. In the case of ERC, the proxy for the oracle is the single-arm policy that always plays the arm i_* with the minimum mean-variance. The performance loss of this optimal single-arm policy $\widehat{\pi}_*$ as compared to π_* is bounded by an $O(N \log T)$ factor (Vakili and Zhao, 2016 [196]):

$$\mathrm{MV}_E(\hat{\pi}_*) - \mathrm{MV}_E(\pi_*) \leq \min\left\{\sigma^2_{\max}\left(\sum_{i\neq i_*}\frac{\Gamma_i^2}{\Delta_i}+1\right), \frac{N}{a}\log T\right\}, \tag{5.20}$$

where $\Delta_i = \mathrm{MV}_i - \mathrm{MV}_{i_*}$, $\Gamma_i = \mu_i - \mu_{i_*}$. The proxy regret has the following form (Vakili and Zhao, 2016 [196]):

$$\hat{R}_\pi(T;\mathbf{F}) = \sum_{i=1}^N(\Delta_i + \Gamma_i^2)\,\mathbb{E}[\tau_i(T)] - \frac{1}{T}\mathbb{E}\left[\left(\sum_{i=1}^N(\hat{\mu}_i(T)-\mu_{i_*})\tau_i(T)\right)^2\right] + \sigma^2_{i_*}, \tag{5.21}$$

where $\sigma^2_{i_*}$ is the variance of arm i_* and $\hat{\mu}_i(T) = \frac{1}{\tau_i(T)}\sum_{s=1}^{\tau_i(T)} X_i(t_s)$ is the empirical mean of the entire realized reward sequence from arm i (t_s denotes the time instant for the sth play of arm i).

Note that under the measure of mean-variance, regret can no longer be written as the sum of certain properly defined immediate performance loss at each time instant. More specifically, under ERC, the contribution from playing a suboptimal arm at a given time t to the overall regret cannot be determined without knowing the entire sequence of decisions and observations. Furthermore, regret in mean-variance involves higher order statistics of the random time $\{\tau_i(T)\}_{i=1}^N$ spent on each arm. Specifically, while the regret in terms of the total expected reward can be written as a weighted sum of the expected value of $\tau_i(T)$ (see (4.6)), regret in terms of mean-variance depends on not only the expected value of $\tau_i(T)$, but also the second moment of $\tau_i(T)$ and the cross correlation between $\tau_i(T)$ and $\tau_j(T)$. These fundamental differences in the behavior of regret are what render the problem difficult and call for different techniques from that used in risk-neutral bandit problems.

Similarly in GRC, the single-arm policy $\widehat{\pi}_*$ is not the optimal policy under known models but can serve as a proxy for the oracle π_* with a similar performance gap (see Vakili and Zhao, 2015 [195]). In LRC, intuitive but not immediately obvious, the oracle π_* is actually the single-arm policy that plays the arm i_* with the minimum mean-variance (Vakili, Boukouvalas, and Zhao, 2019 [193]). Regret expressions under GRC and LRC can be found in Vakili and Zhao, 2015 [195] and Vakili, Boukouvalas, and Zhao, 2019 [193], respectively, which show dependencies on the second-order moments of $\{\tau_i(T)\}_{i=1}^N$ in the case of GRC and on the cumulative variance in the action sequence in the case of LRC. The dependency of regret on the variance of the chosen action sequence under LRC is quite unique. It shows that regret depends not only on statistics of the total times spent on suboptimal arms, the uncertainty (i.e., the cumulative variance) in the action sequence itself is penalized. This indicates that the LRC criterion in certain sense captures the learner's interest in robust decisions and outcomes in a risk-sensitive environment.

Regret lower bounds and order-optimal policies: Establishing regret lower bounds under the mean-variance measures relies on carefully bounding each term in the regret decomposition. Within the uniform-dominance framework, by following a similar line of arguments as in Lai and Robbins, 1985 [126], on the risk-neutral bandit model, the same $\Omega(\log T)$ regret lower bounds can be established under each of the three mean-variance measures. The impact of the risk constraint on regret is absorbed into the leading constants of the $\log T$ order under all three risk measures. In particular, the leading constants have a much stronger dependency on the distribution parameters $\{\Delta_i, \Gamma_i\}_{i=1}^N$ and increase at a much faster rate as $\min_{i=1,\ldots,N} \Delta_i$ diminishes (i.e., the optimal arm i_* in terms of mean-variance becomes harder to identify) while keeping $\{\Gamma_i\}_{i=1}^N$ bounded away from 0.

Such strong dependencies of the leading constants on arm configurations translate to significantly higher minimax regret order which demands a leading constant that holds uniformly for all arm configurations and horizon length T. As a result, the lower bounds on the minimax regret have significantly higher orders than their risk-neutral counterparts. Specifically, the minimax regrets under ERC and LRC are lower bounded by $\Omega(T^{2/3})$ and $\Omega(T)$, respectively, as shown in Vakili and Zhao, 2016 [196] and Vakili, Boukouvalas, and Zhao, 2019 [193]. Of particular significance is that sublinear regret order is no longer possible under LRC in the minimax setting. The lower bound on the minimax regret under GRC remains open.

We now consider risk-averse policies. We present modifications of UCB and DSEE, two representative policies from, respectively, the adaptive and open-loop control families. We focus on the case of ERC; similar variants of UCB and DSEE can be constructed for GRC and LRC.

Referred to as MV-UCB, this variant of UCB-α is similar to that for the risk-neutral bandit (see Algorithm 4.3). The main difference in the index form is that the sample mean is replaced by the sample mean-various $\overline{\mathrm{MV}}_i(t)$ and the policy parameter α needs to be set to different values dependent on the risk tolerance parameter ρ. Specifically, the index of arm i at time $t+1$ (i.e., after t plays) is

$$I_i(t+1) = \overline{\mathrm{MV}}_i(t) - \sqrt{\frac{\alpha \log t}{\tau_i(t)}}, \tag{5.22}$$

and the arm with the *smallest* index is chosen for activation. Note also the minus sign of the second term of the index. This is due to the minimization rather than maximization nature of the mean-variance criterion. Consequently, a lower confidence bound is needed.

The MV-UCB was first introduced and analyzed by Sani, Lazaric, and Munos, 2012 [175], who showed a $\sqrt{T}$ upper bound on the problem-specific regret order of MV-UCB. This bound was later tightened to $\log T$ in Vakili and Zhao, 2016 [196], which, together with the lower bound, demonstrates the order optimality of MV-UCB in the uniform-dominance setting (under the condition that Δ_i is positive for all i). The minimax regret of MV-UCB, however, scales linearly with T.

MV-DSEE is a variation of DSEE by replacing the arm with the largest sample mean by the arm with the smallest sample mean-variance in the exploitation sequence. It achieves the optimal regret orders in both the uniform-dominance setting and the minimax setting by choosing a proper cardinality of the exploration sequence.

Risk neutral vs. risk averse: Table 5.1 compares the regret performance under risk-neutral and risk-averse performance measures. This comparison shows that $R_\pi(T;\mathbf{F})$ has a much stronger dependency on $\mathbf{F}$ under risk measures and achievable minimax regret has significantly higher order. The comparison between adaptive policies such as MV-UCB and open-loop control of exploration and exploitation such as MV-DSEE shows that under risk measures which penalize variations, open-loop policies with less randomness in actions may have an edge over adaptive policies. In particular, under ERC, MV-DSEE achieves the optimal minimax regret order, while MV-UCB incurs linear regret. Under GRC, MV-DSEE preserves its logarithmic problem-specific regret order, while MV-UCB has an order of $\log^2 T$. This can be seen from the dependence of the regret under ERC (see (5.21)) and GRC on the second moments of the random times $\{\tau_i(T)\}_{i=1}^N$. The deterministic nature of DSEE results in deterministic values of $\{\tau_i(T)\}_{i=1}^N$ with zero variance, thus favorable regret order.

Table 5.1: Regret performance under risk-neutral and risk-averse measures

	Problem-Specific Regret		Minimax Regret	
	Lower Bound	Policies	Lower Bound	Policies
Risk Neutral	$\Omega(\log T)$	MV-UCB: $O(\log T)$ MV-DSEE: $O(\log T)$	$\Omega(\sqrt{T})$	MV-UCB: $O(\sqrt{T})$ MV-DSEE: $O(\sqrt{T})$
ERC	$\Omega(\log T)$	MV-UCB: $O(\log \text{ T })$ MV-DSEE: $O(\log \text{ T })$	$\Omega(T^{2/3})$	MV-UCB: $O(T)$ MV-DSEE: $O(T^{2/3})$
LRC	$\Omega(\log T)$	MV-UCB: $O(\log \text{ T })$ MV-DSEE: $O(\log \text{ T })$	$\Omega(T)$	$O(T)$
GRC	$\Omega(\log T)$	MV-UCB: $O(\log^2 \text{ T })$ MV-DSEE: $O(\log \text{ T })$	?	?

Other risk measures: There are also a couple of results on MAB under the measure of *value at risk* (see Galichet, Sebag, and Teytaud, 2013 [83]; Vakili and Zhao, 2015 [195]) and a risk measure based on the logarithm of moment generating function (Maillard, 2013 [141]).

A corresponding expert setting (i.e., non-stochastic formulation with full-information feedback) was studied by Even-Dar, Kearns, and Wortman, 2006 [77], where a negative result was established, showing the infeasibility of sublinear regret under the mean-variance measure.

A related line of studies concerns with the deviation of regret (in terms of cumulative rewards) from its expected value, and high probability bounds have been established by Audibert, Munos, and Szepesvari, 2009 [21] and Salomon and Audibert, 2011 [174].

5.4.2 PURE-EXPLORATION BANDITS: ACTIVE INFERENCE

There is also an inference version of the bandit model with an objective of identifying the top K arms in terms of their mean rewards (without loss of generality, we assume $K = 1$). Under this objective, actions taken during the process are purely for exploration, and rewards generated during the process are merely observations. For this reason, this bandit model is often referred to as pure-exploration bandits.

The problem is one of *active composite hypothesis testing*. It is composite since observations under a given hypothesis (e.g., arm 1 is the best arm) may be drawn from a set of distributions as compared to a single distribution under a simple hypothesis. Specifically, the probabilistic model of the observations under the hypothesis that arm 1 is the best can be any of the arm configurations $\mathbf{F}^N$ satisfying $\mu_1 = \max_{i=1,\ldots,N} \mu_i$. The problem is an active hypothesis testing due to the decision maker's control on where to draw the observations for the inference task by choosing judiciously which arm to pull at each time.

As in classical hypothesis testing, there are two ways to pose the problem: the fixed-sample-size setting and the sequential setting. In the context of pure-exploration bandits, they are often referred to as the fixed-budget and the fixed-confidence settings.

In the fixed-budget setting, the total number of observations/samples is prefixed to T. The objective is to optimize the accuracy of the inference at the end of the time horizon. The inference accuracy can be measured by the probability that the identified best arm $\hat{i}_*$ is indeed the one with the highest mean or by the so-called simple regret given by the gap $\mu_* - \mu_{\hat{i}_*}$ in their mean values. An inference strategy (also referred to as a test) consists of an arm selection rule governing which arm to pull at each time $t = 1, \ldots, T$ and a decision rule governing which arm to declare as the best at the end of the detection process.

In the fixed-confidence setting, the goal is to reach a given level of detection accuracy (e.g., $\Pr[\hat{i}_* \neq i_*] \leq \epsilon$) with a minimum number of samples. The decision horizon is thus random and controlled by the decision maker. An inference strategy consists of, in addition to an arm selection rule and a decision rule as in the fixed-budget setting, a stopping rule that determines when the required detection accuracy is met and no more samples are necessary. Note that the stopping rule is given by a stopping time, not a last-passage time. The latter refers to the time at which the required detection accuracy is met, but the decision maker may not be able to conclude that from past observations. In other words, a sequential inference strategy needs to be able to self terminate.

The general problem of active hypothesis testing was pioneered by Chernoff, 1959 [62], under the term "sequential design of experiments," which was also used in reference to the bandit problem as in Robbins, 1952 [168], Bradt, Johnson, and Karlin, 1956 [43], and Bellman, 1956 [30] in earlier days of the bandit literature. Indeed, to motivate the sequential design problem for testing hypotheses, Chernoff considered the same clinical trial problem posed by Thompson, 1933 [190]. We present in the example below this two-armed bandit problem and give the basic idea of the test developed by Chernoff.

Example 5.2 *A Two-Armed Pure-Exploration Bandit and the Chernoff Test:*
There are two drugs with unknown success probabilities of θ_1 and θ_2, respectively. The objective is to identify, with sufficient accuracy, the drug with higher efficacy using a minimum number of experiments (i.e., the fixed-confidence setting).

Let the two hypotheses be denoted as $H_1 : \theta_1 > \theta_2$ and $H_2 : \theta_1 \leq \theta_2$. Let

$$\Theta_{H_1} = \{\boldsymbol{\theta} = (\theta_1, \theta_2) : \theta_1 > \theta_2\}, \quad \Theta_{H_2} = \{\boldsymbol{\theta} = (\theta_1, \theta_2) : \theta_1 \leq \theta_2\} \tag{5.23}$$

be the parameter sets corresponding to the two hypotheses.

Suppose that we have conducted t experiments, consisting of t_1 trials of drug 1 with m_1 successes and t_2 trials of drug 2 with m_2 successes. We now need to decide which drug to test next, in other words, given what we have observed, whether testing drug 1 or durg 2 would give us more information on which hypothesis is true.

The basic idea proposed by Chernoff for selecting the next experiment is as follows. Given the past t observations, the maximum likelihood estimate of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \left(\frac{m_1}{t_1}, \frac{m_2}{t_2}\right). \tag{5.24}$$

Without loss of generality, suppose that $\frac{m_1}{t_1} > \frac{m_2}{t_2}$. In other words, the maximum likelihood estimate says H_1 is true. An appropriate experiment to take next is one that provides more information for testing the hypothesis $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ vs. the simple alternative $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ where $\tilde{\boldsymbol{\theta}} \in \Theta_{H_2}$ is the parameter point under the alternative hypothesis H_2 that is the most "difficult" to be differentiated from the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. To make this criterion for selecting the next experiment precise, we need to specify a measure for the amount of information provided by the outcome of an experiment and a way for selecting the most "difficult" alternative $\tilde{\boldsymbol{\theta}} \in \Theta_{H_2}$. For the former, a natural candidate is the KL divergence between the two distributions of observations induced by $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, which determines the rate at which the hypothesis $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ (assumed to be the ground truth) can be differentiated from the simple alternative $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. Specifically, Let $D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E)$ denote the corresponding KL divergence under experiment E. For the available two experiments of testing drug 1 ($E = 1$) and testing drug 2 ($E = 2$), we have

$$\begin{aligned} D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E = 1) &= D(\hat{\theta}_1||\tilde{\theta}_1) = \hat{\theta}_1 \log \frac{\hat{\theta}_1}{\tilde{\theta}_1} + (1 - \hat{\theta}_1) \log \frac{1 - \hat{\theta}_1}{1 - \tilde{\theta}_1}, \qquad (5.25) \\ D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E = 2) &= D(\hat{\theta}_2||\tilde{\theta}_2) = \hat{\theta}_2 \log \frac{\hat{\theta}_2}{\tilde{\theta}_2} + (1 - \hat{\theta}_2) \log \frac{1 - \hat{\theta}_2}{1 - \tilde{\theta}_2}, \qquad (5.26) \end{aligned}$$

which follows from the expression of the KL divergence between two Bernoulli distributions. For given $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, the experiment that maximizes $D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E)$ should be selected.

For specifying the most "difficult" parameter point $\tilde{\boldsymbol{\theta}} \in \Theta_{H_2}$, we cast the problem as a zero-sum game with the payoff determined by $D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E)$. The decision maker choose the experiment E to maximize $D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E)$, while nature chooses $\tilde{\boldsymbol{\theta}} \in \Theta_{H_2}$ to minimize $D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E)$. The solution for the decision maker to this zero-sum game is a mixed strategy that chooses $E = 1$ with probability p_* and $E = 2$ with probability $1 - p_*$ where p_* is determined by the following maximin problem:

$$\begin{aligned} p_* &= \arg \max_{p \in [0,1]} \min_{\tilde{\boldsymbol{\theta}} \in \Theta_{H_2}} \Big(pD(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E = 1) + (1 - p)D(\hat{\boldsymbol{\theta}}||\tilde{\boldsymbol{\theta}}; E = 2) \Big) \qquad (5.27) \\ &= \arg \max_{p \in [0,1]} \min_{\tilde{\boldsymbol{\theta}} \in \Theta_{H_2}} \Big(pD(\hat{\theta}_1||\tilde{\theta}_1) + (1 - p)D(\hat{\theta}_2||\tilde{\theta}_2) \Big). \qquad (5.28) \end{aligned}$$

The above specifies the selection rule for choosing experiments. The stopping rule is as follows. Suppose again that the current maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\frac{m_1}{t_1}, \frac{m_2}{t_2})$ satisfies $\frac{m_1}{t_1} > \frac{m_2}{t_2}$. The maximum likelihood estimate $\check{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ under the alternative hypothesis H_2 is given by

$$\check{\boldsymbol{\theta}} = \left(\frac{m_1 + m_2}{t_1 + t_2}, \frac{m_1 + m_2}{t_1 + t_2}\right). \tag{5.29}$$

Note that the maximum likelihood estimate of $\boldsymbol{\theta}$ under H_2 is the parameter point in Θ_{H_2} that is most likely to generate the given observations of m_1 successes from t_1 trials of drug 1 and m_2 successes from t_2 trials of drug 2, which would be the point of $\theta_1 = \theta_2$ for the observations satisfying $\frac{m_1}{t_1} > \frac{m_2}{t_2}$. The stopping rule is then to terminate the test when the observations give sufficient evidence to differentiate the most likely point $\hat{\boldsymbol{\theta}}$ under H_1 from the most likely point $\check{\boldsymbol{\theta}}$ under H_2. Specifically, the termination condition is that the sum of the log-likelihood ratio of $\hat{\boldsymbol{\theta}}$ to $\check{\boldsymbol{\theta}}$ exceeds a threshold determined by the required inference accuracy:

$$m_1 \log \frac{\hat{\theta}_1}{\check{\theta}_1} + (t_1 - m_1) \log \frac{1 - \hat{\theta}_1}{1 - \check{\theta}_1} + m_2 \log \frac{\hat{\theta}_2}{\check{\theta}_2} + (t_2 - m_2) \log \frac{1 - \hat{\theta}_2}{1 - \check{\theta}_2} > -\log c, \tag{5.30}$$

where c is the cost for each sample/experiment, which can be intepretated as the inverse of the Lagrange multiplier for the constraint on the detection error.

The decision rule at the time of stopping is to simply declare the hypothesis corresponding to the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ (i.e., declare H_1 if $\frac{m_1}{t_1} > \frac{m_2}{t_2}$ and H_2 otherwise). An implementation of the above Chernoff test is given in Algorithm 5.6.

The above Chernoff test applies to general active composite hypothesis testing problems with general distributions and an arbitrary finite number of experiments, but involving only two hypotheses and a finite number of states of nature. Chernoff, 1959 [62] showed that this test is asymptotically optimal in terms of its Bayes risk as c—equivalently, the maximum allowed probability of detection error ϵ—tends to zero.

Bessler, 1960 [38] extended Chernoff's result to cases with an arbitrary finite number of hypotheses and a potentially infinite number of experiments. Albert, 1961 [10] considered a more complex composite model for the hypotheses, allowing an infinite number of states of nature. Recent results on variations and extensions include Nitinawarat, Atia, Veeravalli, 2013 [161] (where the problem was referred to as controlled sensing for hypothesis testing) and Naghshvar and Javidi, 2013 [151].

The pure-exploration bandit problem is a special case of active composite hypothesis testing. In particular, in a general active hypothesis testing problem, the number of hypothesis can be independent of the number of experiments, while in pure-exploration bandits, these two are equal and are given by the number of arms. More importantly, the observation distributions can be arbitrary across hypotheses and available experiments in a general active hypothesis testing problem. In pure-exploration bandits, however, the observation distributions for the N^2 hypothesis-action pairs are coupled through a fixed set of N reward distributions. Representative studies on the pure-exploration bandits include Mannor and Tsitsiklis, 2004 [145], Even-Dar, Mannor, Mansour, 2006 [78], Audibert, Bubeck, and Munos, 2010 [22], and Bubeck, Munos,

Algorithm 5.6 The Chernoff Test for a Pure-Exploration Two-Arm Bernoulli Bandit

Notations: (m_i, t_i): the number of successes and the number of plays of arm i up to the current time;
Input: c: the cost of each observation/experiment.

```
Initialization:
Pull each arm once in the first two plays.
Set TERMINATE = 0.
while TERMINATE = 0 do
    Obtain the maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\frac{m_1}{t_1}, \frac{m_2}{t_2})$.
    Obtain the maximum likelihood estimate under the alternative hypothesis: $\check{\boldsymbol{\theta}} = (\frac{m_1+m_2}{t_1+t_2}, \frac{m_1+m_2}{t_1+t_2})$.
    if the termination condition in (5.30) is met then
        TERMINATE = 1.
        Declare the hypothesis corresponding to $\hat{\boldsymbol{\theta}}$.
    else
        Compute the probability $p_*$ as given in (5.28) (replacing $\Theta_{H_2}$ with $\Theta_{H_1}$ if $\frac{m_1}{t_1} \leq \frac{m_2}{t_2}$).
        Play arm 1 with probability $p_*$ and arm 2 with probability $1 - p_*$.
        Update $(m_i, t_i)$ based on the chosen action and the resulting outcome.
    end if
end while
```

Stoltz, 2011 [47]. A variation of this model is the thresholding bandits where the objective is to identify arms with mean values above a given threshold (Locatelli, Gutzeit, and Carpentier, 2016 [138]). We see in Section 6.1.3 a variant of the thresholding bandits that allows arm aggregation in the application of heavy hitter detection in Internet traffic engineering and security.

5.5 LEARNING IN CONTEXT: BANDITS WITH SIDE INFORMATION

Consider again the example of clinical trial. Suppose that the patient population consists of two types. For the first type of patients, which makes up 80% of the population, the efficacy of the two drugs is given by $\theta_1^{(1)} = 0.7$ and $\theta_2^{(1)} = 0.4$. For the second type, $\theta_1^{(2)} = 0.2$, $\theta_2^{(2)} = 0.7$. At each time t, we face a patient chosen uniformly at random from the population. If the type of the patient is not observable, the problem is the classic bandit model we have discussed with the unknown mean rewards of the two arms given by $\mu_1 = 0.8\theta_1^{(1)} + 0.2\theta_1^{(2)} = 0.6$ and

$\mu_2 = 0.8\theta_2^{(1)} + 0.2\theta_2^{(2)} = 0.46$. The optimal arm here is arm 1, and online learning algorithms discussed in Chapter 4 would converge to the optimal average performance of arm 1 and cure 60% of patients by prescribing the first drug to all but a $\log T$ order of patients. Type 2 patients, however, are prescribed with the inferior drug most of the time and experience a success rate of only 40%.

It is easy to see that if the type of each patient is known, better performance can be achieved. The problem can be simply decomposed into two independent bandit problems interleaved in time, one for each type of patients. For each type of patients, the better drug (drug 1 for type 1 and drug 2 for type 2) would be prescribed most of the time, achieving the optimal average performance of 0.7 for each type.

This is the rationale behind the contextual bandit model, also referred to as bandits with side information and bandits with covariates. Motivating applications include personalized medicine, personalized advertizement and product/content recommendations, where certain user features (referred to as the context) may be known.

Definition 5.3 *The Contextual Bandit Model:*
At each time t, a context Y is drawn, i.i.d. over time, from an unknown distribution $F_Y(y)$ and revealed to the player. In context $Y = y$, rewards from arm i $(i = 1, \ldots, N)$ obey an unknown distribution $F_i(x|y)$. Rewards are independent over time and across arms. Let $\mu_i(y)$ denote the expected reward of arm i in context y. Define

$$\mu_*(y) \stackrel{\Delta}{=} \max_{i=1,\ldots,N} \mu_i(y), \quad i_*(y) \stackrel{\Delta}{=} \arg\max_{i=1,\ldots,N} \mu_i(y) \tag{5.31}$$

as, respectively, the maximum expected reward and the optimal arm in context y. The regret of policy π over T plays is given by

$$R_\pi(T) = \sum_{t=1}^{T} \mathbb{E}_Y \left[\mu_*(Y) - \mu_{\pi(t)}(Y)\right], \tag{5.32}$$

where $\pi(t)$ denotes the arm chosen by π at time t.

While the gain in performance from such contextual information can be significant, the problem does not offer much intellectually if it simply decomposes into independent bandit problems based on the context. Interesting formulations arise when certain coupling across these bandit problems in different contexts is present and can be leveraged in learning.

Three coupling models have been considered in the literature. The first model is parametric, assuming that the reward distributions of arm i in different contexts are coupled through a common unknown parameter θ_i. Specifically, pulling arm i in context $y \in \mathcal{Y}$ generates random rewards according to a known distribution $F_i(x; \theta_i \mid y)$ with an unknown parameter θ_i (see Wang, Kulkarni, and Poor, 2005 [205]). For instance, for a binary context space $\mathcal{Y} = \{-1, 1\}$, the reward distribution of arm i under each context $y \in \{-1, 1\}$ is Gaussian with mean $y\theta_i$ and variance 1. Another example is a linear parametric coupling in the mean rewards of arms in different contexts (see Li et al., 2010 [131] and Chu et al., 2011 [64]). Specifically, the context is assumed to be a d-dimensional feature vector $\mathbf{y}$, and the expected reward of arm i in context $\mathbf{y}$ is given by $\mu_i(\mathbf{y}) = \mathbf{y}^T \boldsymbol{\theta}_i$ for an unknown parameter vector $\boldsymbol{\theta}_i$ of arm i. Under this model, observations obtained in all contexts can be used collectively to infer the underlying common parameters, which then guides the arm selection tailored specifically to each context.

The second model stems from viewing the problem as an online reward-sensitive multi-class classification where the context y at each time is the instance and the arm pulling action equates to assigning a label $i \in \{1, \ldots, N\}$ to y. The most rewarding label of y is $i_*(y)$. A commonly adopted formulation in such supervised learning problems is to specify a hypothesis set consisting of all classifiers under consideration, and the objective is to learn the best classifier in this set. In the context of contextual bandits, the hypothesis set specifies mappings from contexts to arms. Regret is then defined with respect to the best mapping in this set (Langford and Zhang, 2007 [128] and Seldin et al., 2011 [176]). If the hypothesis set includes the optimal mapping $y \to i_*(y)$, the corresponding regret is the same as defined in (5.32). The inherent structure of the hypothesis set couples the different bandit problems across contexts. Consider a rather extreme example: if all mappings in the hypothesis set map contexts y_1 and y_2 to an identical arm, then bandit problems corresponding to these two contexts can be aggregated as a single learning problem. On the other hand, if the hypothesis set contains all possible mappings from $\mathcal{Y}$ to $\{1, \ldots, N\}$, then no structure exists, and the contextual bandit problem decomposes into independent bandit problems across contexts.

In the third model, the context is assumed to take continuum values (or in a general metric space). The mean reward $\mu_i(y)$ of arm i is assumed to be a Lipschitz continuous function of the context y. In other words, each arm yields similar rewards (in expectation) in similar contexts. This naturally leads to an approach that discretizes the context space $\mathcal{Y}$ and aggregates contexts in the same discrete bin as a single bandit problem while decoupling contexts across different bins as independent bandit problems (see Rigollet and Zeevi, 2010 [167]). The zooming algorithm developed by Kleinberg, Slivkins, and Upfal, 2015 [117] for Lipschitz bandits has been extended to contextual bandit problems where the Cartesian product of the context space and the arm space belongs to a general metric space and the mean reward $\mu(a, y)$ as a function of the arm a and the context y is Lipschitz (Slivkins, 2014 [178]). As discussed in Section 5.2.1, the problem will be more interesting if, instead of an open-loop discretization tailored to the worst-case smoothness condition, the objective is set to a discretization strategy that automatically adapts to the local smoothness of the unknown mean reward functions $\{\mu_i(y)\}_{i=1}^N$.

5.6 LEARNING UNDER COMPETITION: BANDITS WITH MULTIPLE PLAYERS

The canonical bandit model and the variants discussed above all assume a single player. Many applications give rise to a bandit problem with multiple players. What couples these players together is the contention among players who pull the same arm. When multiple players choose the same arm, the reward offered by the arm is distributed arbitrarily among the players, not necessarily with conservation. Such an event is referred to as a collision.

Depending on whether each player has access to other players' past actions and observations, the problem has been addressed under two settings: the centralized and the distributed settings.

5.6.1 CENTRALIZED LEARNING

Consider a bandit model with N arms and K players with $K < N$. When multiple players pull the same arm, no reward is accrued from this arm by any player. Extensions to other collision models where one player captures the reward or players share the reward in a certain way are often straightforward. We assume that the expected reward is nonnegative for all arms. Collisions are thus undesired events. Regret is defined in terms of the system-level reward—the total reward summed over all players.

In the centralized setting, each player has access to all other players' past actions and observations, and can thus coordinate their actions and act collectively as a single player. The problem can thus be treated as a bandit model with a single player who pulls K arms simultaneously at each time. Referred to as bandits with multiple plays, this model was studied by Anantharam, Varaiya, and Walrand, 1987a [15] under the uniform-dominance approach, where the technical treatment follows that of Lai and Robbins, 1985 [126].

A more complex model is when each arm offers different rewards when pulled by different players. This scenario arises in applications such as multichannel dynamic access in wireless networks where users experience different channel qualities due to, for example, differences in their geographical locations. Another application example is multi-commodity bidding where a commodity may have different values to different bidders.

Let μ_{ij} denote the expected reward offered by arm i when pulled by player j. Under the known model, the oracle policy is given by a maximum bipartite matching. Specifically, consider a bipartite graph with two sets of vertices representing, respectively, the set of N arms and the set of K players. An edge exists from arm i to player j with weight μ_{ij}. A maximum matching on this bipartite graph gives the optimal matching of the K players to the N arms and leads to the maximum sum reward that defines the benchmark. When the model is unknown, the problem is a special case of the combinatorial semi-bandit model discussed in Section 5.2.1. Specifically, using the graphical random field model as illustrated in Figure 5.1, each matching ϕ of the K players to N arms constitutes a super-arm X_ϕ. The latent variables are the random reward $Y_{i,j}$ from arm i pulled by player j. Let $\phi(j)$ denote the arm assigned to player j under matching ϕ.

Then

$$X_{\phi} = \sum_{j=1}^{K} Y_{\phi(j),j}, \tag{5.33}$$

where each $Y_{\phi(j),j}$ is observed when matching ϕ is engaged (i.e., the semi-bandit feedback).

A UCB-based learning policy was developed by Gai, Krishnamachari, and Jain, 2011 [82], where a bipartite matching algorithm is carried out at each time. The unknown edge weights $\{\mu_{i,j}\}_{i=1,\ldots,N; j=1,\ldots,K}$ of the bipartite graph of arms and players is replaced by the UCB-α indexes calculated from past observations. A matching algorithm (e.g., the Hungarian algorithm [122]) is then carried out on this bipartite graph to find the best matching, which determines the action of each player at next time. This algorithm was shown to achieve the optimal logarithmic problem-specific regret order.

5.6.2 DISTRIBUTED LEARNING

In the distributed setting, each player, not able to observe other players' actions or rewards, decides which arm to pull based on its own local observations. A decentralized learning policy π is given by the concatenation of the local polices for all players: $\pi = (\pi_1, \ldots, \pi_K)$. The performance benchmark is set to be the same as in the centralized setting: the oracle knows the reward model and carries out a centralized optimal matching of players to arms that eliminates collisions among players.

Under a distributed learning algorithm, however, collisions are inevitable. Collisions reduce not only the system-level reward, but also available observations for learning. If collisions are unobservable, then observations are contaminated: players do not know whether the observed reward truly reflect the quality of the chosen arm. Nevertheless, when the reward mean of each arm is identical across players, the optimal logarithmic regret order can be achieved as in the centralized setting. Several distributed learning policies with the optimal regret order have been developed in various settings regarding fairness among players and the observability of collision events (see, for example, Liu and Zhao, 2010 [136]; Anandkumar et al., 2011 [14]; Vakili et al., 2013 [194]; Gai and Krishnamachari, 2014 [81]; Rosenski, Shamir, and Szlak, 2016 [169]; and references therein).

In the case where each arm offers different rewards when pulled by different players, the problem is more complex. Kalathil, Nayyar, and Jain, 2014 [105], developed a learning policy that integrates a distributed bipartite matching algorithm (e.g., the auction algorithm by Bertsekas, 1992 [34]) with the UCB-α policy and showed that it achieves $O(\log^2 T)$ problem-specific regret order. Recently, Bistritz and Leshem, 2018 [39] closed the gap and developed a policy with a near-$O(\log T)$ regret order. This policy leverages recent results on a payoff-based decentralized learning rule for a generic repeated game by Pradelski and Young, 2012 [163] and Marden, Yong, and Pao, 2014 [146]. The problem is also recently studied under the non-stochastic setting by Bubeck et al., 2019 [51].

CHAPTER 6

Application Examples

We consider in this chapter a couple of representative application examples in communication networks and social-economical systems. We hope to illustrate not only the general applicability of the various bandit models, but also the connections and differences between the Bayesian and the frequentist approaches. Our focus will be on the bandit formulations of these application examples. Key ideas are highlighted with details of the specific solutions left to the provided references.

6.1 COMMUNICATION AND COMPUTER NETWORKS

6.1.1 DYNAMIC MULTICHANNEL ACCESS

Consider the problem of probing N independent Markov chains. Each chain has two states—good (1) and bad (0)—with transition probabilities $\{p_{01}, p_{11}\}$. At each time, a player chooses K $(1 \leq K < N)$ chains to probe and receives a reward for each probed chain that is in the good state. The objective is to design an optimal policy that governs the selection of K chains at each time to maximize the long-run reward.

The above general problem is archetypal of dynamic multichannel access in communication systems, including cognitive radio networks, downlink scheduling in cellular systems, opportunistic transmission over fading channels, and resource-constrained jamming and anti-jamming. In the communications context, the N independent Markov chains correspond to N communication channels under the Gilbert-Elliot channel model (Gilbert, 1960 [86]), which has been commonly used to abstract physical channels with memory. The state of a channel models the communication quality and determines the reward of accessing this channel. For example, in cognitive radio networks where secondary users search in the spectrum for idle channels temporarily unused by primary users (Zhao and Sadler, 2007 [219]), the state of a channel models the occupancy of the channel by primary users. For downlink scheduling in cellular systems, the user is a base station, and each channel is associated with a downlink mobile receiver. Downlink receiver scheduling is thus equivalent to channel selection. The application of this problem also goes beyond communication systems. For example, it has applications in target tracking as considered by Le Ny, Dahleh, and Feron, 2008 [130], where K unmanned aerial vehicles are tracking the states of N $(N > K)$ targets in each slot.

Known Markovian dynamics—a restless bandit problem: Assume first that the Markov transition probabilities $\{p_{01}, p_{11}\}$ are known. We show that in this case, the problem is a restless multi-armed bandit problem discussed in Section 3.3. Without loss of generality, we focus on the case of $K = 1$.

Let $[S_1(t), \ldots, S_N(t)] \in \{0, 1\}^N$ denote the channel states in slot t. They are not directly observable before the sensing action is made. The user can, however, infer the channel states from its decision and observation history. A sufficient statistic for optimal decision making is given by the conditional probability that each channel is in state 1 given all past decisions and observations (see, e.g., Smallwood and Sondik, 1971 [181]). Referred to as the belief vector or information state, this sufficient statistic is denoted by $\boldsymbol{\omega}(t) \triangleq [\omega_1(t), \cdots, \omega_N(t)]$, where $\omega_i(t)$ is the conditional probability that $S_i(t) = 1$. Given the sensing action a and the observation in slot t, the belief state in slot $t + 1$ can be obtained recursively as follows:

$$\omega_i(t+1) = \begin{cases} p_{11}, & \text{if } i = a \text{ and } S_i(t) = 1 \\ p_{01}, & \text{if } i = a \text{ and } S_i(t) = 0 \\ \Phi(\omega_i(t)), & \text{if } i \neq a \end{cases}, \tag{6.1}$$

where

$$\Phi(\omega_i(t)) \triangleq \omega_i(t) p_{11} + (1 - \omega_i(t)) p_{01}$$

denotes the operator for the one-step belief update for unobserved channels. The initial belief vector $\omega(1)$ can be set to the stationary distribution of the underlying Markov chain, if no information on the initial system state is available.

It is now easy to see that the problem is a restless bandit problem, where each channel constitutes an arm and the state of arm i in slot t is the belief state $\omega_i(t)$. The user chooses an arm a to activate (sense), while other arms are made passive (unobserved). The active arm offers an expected reward of $R_a(\omega_a(t)) = \omega_a(t)$, depending on the state $\omega_a(t)$ of the chosen arm. The states of active and passive arms change as given in (6.1). The performance measure can be set to either total discounted reward or average reward. The former, besides capturing delay-sensitive scenarios, also applies when the horizon length is a geometrically distributed random variable. For example, a communication session may end at a random time, and the user aims to maximize the number of packets delivered before the session ends. The latter is the common measure of throughput in the context of communications.

For this special case of restless bandits, the underlying two-state Markov chain that governs the state transition of each arm brings rich structures into the problem, leading to positive outcomes on the indexability and optimality of the Whittle index policy. Specifically, the indexability and the closed-form expression of the Whittle index under the total-discounted-reward criterion were established independently and using different techniques by Le Ny, Dahleh, and Feron, 2008 [130] and Liu and Zhao, 2008 [134]. The semi-universal structure and the optimality of the Whittle index policy and the analysis under the average-reward criterion were given by Liu and Zhao, 2010 [135].

In particular, for this restless bandit problem, the Whittle index policy is optimal for all N and K in the case of positively correlated channels. For negatively correlated channels, the

Whittle index policy is optimal for $K = 2$ and $K = N - 1$. For a general K, bounds on the approximation ratio of the Whittle index policy were given by Liu and Zhao, 2010 [135]. The optimality of the Whittle index policy follows directly from its equivalence to the myopic policy and the optimality of the myopic policy established by Ahmad et al., 2009 [9].

The semi-universal structure of the Whittle index policy renders explicit calculation of the Whittle index unnecessary when implementing the policy (note that it is the ranking not the specific values of the arm indices that determine the actions). Actions can be determined by maintaining a simple queue with no computation. Specifically, all N channels are ordered in a queue, with the initial ordering $\mathcal{K}(1)$ given by a decreasing order of their initial belief values. In each slot, those K channels at the head of the queue are sensed. Based on the sensing outcomes, channels are reordered at the end of each slot according to the following simple rules (see Figure 6.1).

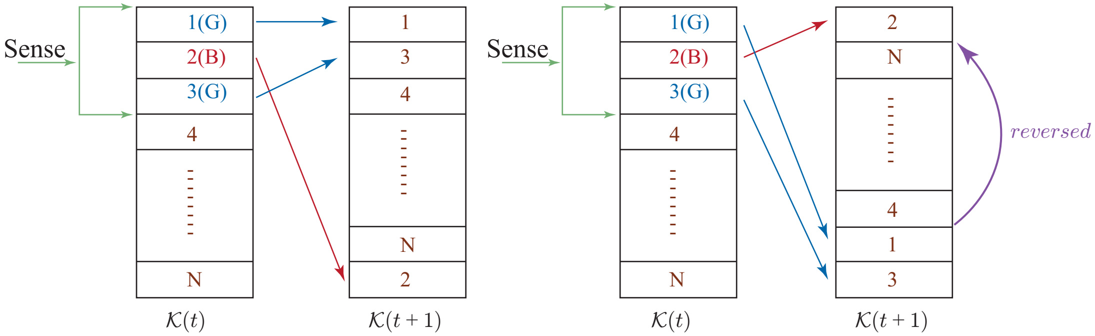


Figure 6.1: The semin-universal structure of the Whittle index policy (the left sub-figure is for $p_{11} \geq p_{01}$, the right for $p_{11} < p_{01}$; G indicates state 1 (good) and B state 0 (bad)) [135]. Used with permission.

In the case of $p_{11} \geq p_{01}$, the channels observed in state 1 will stay at the head of the queue while the channels observed in state 0 will be moved to the end of the queue.

In the case of $p_{11} < p_{01}$, the channels observed in state 0 will stay at the head of the queue while the channels observed in state 1 will be moved to the end of the queue. The order of the unobserved channels are reversed.

The above structure of the Whittle index policy was established based on the monotonicity of the index with the belief value ω, which implies an equivalence between the Whittle index policy and the myopic policy that activates the arm with the greatest belief value (hence the greatest immediate reward). The semi-universal structure thus follows from that of the myopic policy established by Zhao, Krishnamachari, and Liu, 2008 [218]. This structure is intuitively appealing. In the case of $p_{11} > p_{01}$, the channel states in two consecutive slots are positively correlated; a good state is more likely to be followed by a good state than a transition to the bad state. The action is thus to follow the winner. In the case of $p_{11} < p_{01}$, the states in two consecutive slots are negatively correlated. The policy thus stays with the channel currently in the bad state and flips the order of unobserved channels.

In addition to computation and memory efficiency, a more significant implication of this semi-universal structure is that it obviates the need for knowing the channel transition probabilities except the order of p_{11} and p_{01}. As a result, Whittle index policy is robust against model mismatch and automatically tracks variations in the channel model provided that the order of p_{11} and p_{01} remains unchanged. In the case when there is no knowledge on p_{11} and p_{01}, the resulting frequentist bandit model with restless Markov reward processes has a much smaller learning space to explore as compared to the general case discussed in Section 5.1.2. We explore this next.

Unknown Markovian dynamics—frequentist bandit with restless Markov reward processes: When the channel transition probabilities are unknown, the problem can be formulated within the frequentist framework with restless Markovian reward processes as discussed in Section 5.1.2.

The semi-universal structure and the strong performance of the Whittle index policy make it a perfect proxy for the oracle. Specifically, the Whittle index policy follows one of the two possible strategies depending on the order of p_{11} and p_{01}. These two strategies can thus be considered as two meta-arms in a bandit model where the objective is to learn which meta-arm is more rewarding for the underlying unknown Markov dynamics.

A key question in this meta-learning problem is how long to operate each meta-arm at each step. Since each meta-arm is an arm selection strategy in the original N-arm bandit model, it needs to be employed for a sufficiently long period in order to gather sufficient observations on the mean reward it offers. On the other hand, staying with one meta-arm—which can be the suboptimal one—for too long may leads to poor regret performance. A solution given in Dai et al., 2011 [71] is to increase the duration of each activation of the meta-arms at an arbitrarily slow rate. The total reward obtained during each activation (of multiple consecutive time instants) is used to update an UCB-type of index for each meta-arm, which guides the selection of the next meta-arm. It was shown that a regret order arbitrarily close to the optimal logarithmic order can be achieved, with the optimality gap determined by the rate of the diverging sequence for meta-arm activations which can be chosen based on the performance requirement of the application at hand.

6.1.2 ADAPTIVE ROUTING UNDER UNKNOWN LINK STATES

Consider a communication network represented by a directed graph consisting of all simple paths between a given source-destination pair. Let N and M denote, respectively, the number of paths from the source to the destination and the number of edges involved.

At each time t, a random weight/cost $W_i(t)$ drawn from an unknown distribution is assigned to each edge ($i = 1, \ldots, M$). We assume that $W_i(t)$ is i.i.d. over time for all i. At the beginning of each time slot t, a path p is chosen for a message transmission. Subsequently, depending on the observation model, either the random cost of each link in the chosen path is revealed (the semi-bandit feedback) or only the total end-to-end cost, given by the sum of the costs of all links on the path, is observed. The problem is a combinatorial bandit with N arms dependent on M latent variables (see Figure 5.1).

Let $C(p)$ denote the total end-to-end cost of path p, and p_* the optimal route with the minimum expected cost. The regret of an adaptive routing strategy π is given by

$$R_\pi(T) = \mathbb{E}\left[\sum_{t=1}^{T}(C(p_{\pi(t)}) - C(p_*))\right]. \tag{6.2}$$

The objective is a learning policy that offers good regret scaling in M, the number of unknowns, rather than N, the number of arms, while preserving the optimal order in T.

Under the link-level observation model, a UCB-α index can be maintained for each link based on link cost observations, and a shortest path algorithm using the UCB-α indexes as the edge weights can be carried out for path selection at each time (Gai et al., 2011 [82]). The more complex path-level observation model can be handled through a dimension-reduction approach that constructs a barycentric spanner of the underlying graph. The online policy subsequently explores only paths in the barycentric spanner and uses the observed costs of these basis paths to estimate the costs of all paths for exploitation. This approach was first proposed by Awerbuch and Kleinberg, 2008 [27], under the non-stochastic/adversarial setting and later explored in the stochastic setting by Liu and Zhao, 2012 [137]. Such a dimension reduction approach also applies to other combinatorial bandit problems under the strict bandit feedback model.

6.1.3 HEAVY HITTER AND HIERARCHICAL HEAVY HITTER DETECTION

In Internet and other communication and financial networks, it is a common observation that a small number of flows, referred to as heavy hitters (HH), account for most of the total traffic. Quickly identifying the heavy hitters (flows with a packet count per unit time exceeds a given threshold) is thus crucial to network stability and security. In particular, heavy hitter detection is important for a variety of network troubleshooting scenarios, including detecting denial-of-service (DoS) attacks.

The challenge of the problem is that the total number N of flows seen by a router/switch is much larger than the available sampling resources. Maintaining a packet count of each individual flow is infeasible. The key to an efficient solution is to consider prefix aggregation based on the source or destination IP addresses. This naturally leads to a binary tree structure (as illustrated in Figure 6.2) with each node representing an aggregated flow with a specific IP prefix. The packet count of each node on the tree equals to the sum of the packet counts of its children.

A more complex version of the problem is hierarchical heavy hitter (HHH) detection, in which the search for flows with abnormal volume extends to aggregated flows in the upper levels of the IP-prefix tree. Specifically, an upper-level node is an HHH if its mean remains

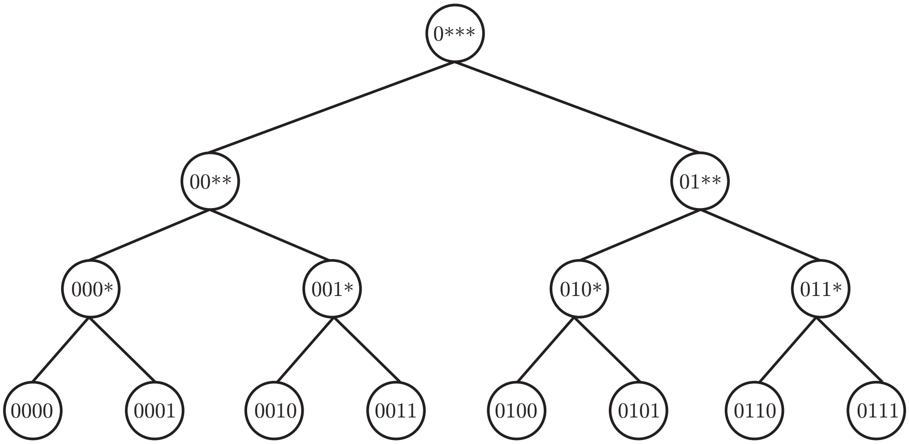


Figure 6.2: An IP-prefix tree for heavy hitter detection.

above a given threshold after excluding all its abnormal descendants (if any). HHH detection is of particular interest in detecting *distributed* denial-of-service attacks that split total traffic over multiple routes.

The problem of detecting heavy hitters and hierarchical heavy hitters can be viewed as a variant of the thresholding bandit with the option of aggregating subset of arms conforming to a given tree structure. An active inference strategy determines, sequentially, which node on the tree to probe and when to terminate the search in order to minimize the sample complexity for a given level of detection reliability. A question of particular interest is how to achieve an optimal sublinear scaling of the sample complexity with respect to the number of traffic flows.

Vakili et al., 2018 [197], proposed an active inference strategy that induces a biased random walk on the tree based on confidence bounds of sample statistics. Specifically, the algorithm induces a biased random walk that initiates at the root of the tree and eventually arrives and terminates at a target with the required reliability. Each move in the random walk is guided by the output of a local confidence-bound based sequential test carried on each child of the node currently being visited by the random walk. The sequential test draws samples from a child sequentially until either a properly designed upper confidence bound drops below a threshold (which results in a negative test outcome indicating this node is unlikely to be an ancestor of a target) or a lower confidence bound exceeds a threshold (resulting in a positive test outcome indicating this node is likely to be an ancestor of a target). The next move of the random walk is then determined based on the test outcome: move to the (first) child tested positive and, if both children tested negative, move back to the parent. The confidence level of the local sequential test is set to ensure that the random walk is more likely to move toward the target than move away from it and that the random walk terminates at a true target with a sufficiently high probability. The sample complexity of this algorithm was shown to achieve the optimal logarithmic order in the number of flows and the optimal order in the required detection accuracy (Vakili et al., 2018 [197]).

6.2 SOCIAL-ECONOMIC NETWORKS

6.2.1 DYNAMIC PRICING AND THE PURSUIT OF COMPLETE LEARNING

The connection between dynamic pricing and the multi-armed bandit model was first explored by Rothschild in 1974 [172], where he constructed an archetype of rational and optimizing sellers facing unknown demand functions of their customers. Rothschid suggested that "A firm which does not know the consequences of charging a particular price has an obvious way of finding out. It may charge that price and observe the result." That is the basic premise of bandit problems.

Dynamic pricing as a Bayesian bandit problem: In a highly stylized model, Rothschild assumed that a firm, trying to price a particular good, sequentially chooses one of two possible prices p_1 and p_2 to a stream of potential customers and observes either success or failure in each sale attempt. The characteristic of each customer is assumed to be probabilistically identical. More specifically, from the point of view of the firm, each potential customer is a Bernoulli random variable with parameter θ_1 if price p_1 is offered and θ_2 if price p_2 is offered. The expected profit when price p_i $(i = 1, 2)$ is offered is $\theta_i(p_i - c)$, where c is the marginal cost of one unit of the good. The objective of the firm is a policy for dynamically setting the price to maximize its long-run cumulative profit.

This is the same two-armed bandit problem as posed by Thompson, 1933 [190] by likening the two prices as the two treatments in a clinical trail. Rothschild, too, took the Bayesian approach, assuming prior knowledge on θ_1 and θ_2 in the form of their prior distributions over $[0, 1]$. He, however, adopted the total-discounted-reward criterion as in the canonical Bayesian bandit model defined in Definition 2.2. The main message of Rothschild's work is that *incomplete learning* occurs with positive probability. Specifically, with probability one, the seller charges only one price infinitely often after a finite duration of price experimentation, and which price the seller eventually settles on depends on those finite sequence of random sale outcomes during the price experimentation. Since the same finite sequence of outcomes can occur with positive probability under both scenarios regarding which of the two arms is more profitable, with positive probability the seller settles on a suboptimal price and terminates learning.

The result is perhaps not surprising for the following reasons. First, the geometric discounting quickly diminishes the future benefit of an accurate learning of the best arm. Second, the objective under the Bayesian framework is the expected performance averaged over all possible values of θ_1 and θ_2 under the known prior distributions rather than insisting good performance under every realization of θ_1 and θ_2. The optimal strategy is thus to accept the possibility of settling on the suboptimal arm caused by small probability events in order to latch on the optimal arm quickly in most cases. Rothschild suggested that such an outcome of incomplete learning under the optimal policy might explain the pervasive phenomenon of price dispersion in a market where the same good is being sold at a variety of prices: sellers, each behaving optimally, settle on different prices due to differences in their realized sale outcomes during price experimentation.

A major follow-up work was given by McLennan, 1984 [148], where he showed that incomplete learning can occur even when the seller has a continuum of price options. The continuum of prices are, however, bound by a specific demand function that is known to take one of two possible forms. Specifically, the underlying demand model is either $\rho_1(p)$ or $\rho_2(p)$, which characterizes the probability of a successful sale at each possible price p. The two-price model considered by Rothschild can be viewed under this model by McLennan by restricting offered prices to the two optimal prices under each possible demand model:

$$p_i = \arg\max_p \rho_i(p)(p - c). \tag{6.3}$$

The model by Rothschild, however, does not assume the knowledge of the function form of the underlying demand curve. Even though this does not alter the conclusion on incomplete learning within the Bayesian framework, it leads to drastically different outcomes within the frequentist framework as discussed below.

Dynamic pricing as a frequentist bandit problem: The fact that incomplete learning occurs with positive probability even under the optimal pricing strategy can be unsettling from the point of view of an individual seller. Seeing only one realization of the demand model and caring about only its own revenue rather than the ensemble average over a large population of probabilistically identical sellers, an individual seller might be more interested in a policy that grantees complete learning under every possible demand model, even though it may not offer the best ensemble-average performance. This brings us to the frequentist formulation of the problem.

Take the market model considered by Rothschild. Under the frequentist formulation, the regret grows unboundedly, either at a logarithmic rate under the uniform-dominance setting or a $\sqrt{T}$ rate under the minimax setting. This implies that the optimal strategy, guarding against every possibility of settling on the suboptimal price, never settles on either price and experimenting each price infinitely often (although with vanishing frequency). If we define complete learning as a probability-one convergence to the optimal price, learning within the frequentist framework is incomplete with probability one, not just positive probability as in the Bayesian framework.

If we take the model by McLennan, however, the conclusion is drastically different. Complete learning can be achieved with probability one under the frequentist framework as shown in Zhai et al., 2011 [217] and Tehrani, Zhai, and Zhao, 2012 [188]. Even if we restrict our actions to the two optimal prices p_1 and p_2 under the two possible demand curves, the knowledge on the specific function forms of the demand curves results in a completely different two-arm bandit problem from that under the Rothschild model. Specifically, since the two arms p_1 and p_2 share the same underlying demand curve, observations from one arm also provide information on the quality of the other arm by revealing the underlying demand curve. Recognizing the detection component of this bandit problem, Zhai et al. proposed a policy based on the likelihood ratio test (LRT): at each time t, the likelihood ratio of seeing the past $t-1$ sale outcomes under demand curve ρ_1 to that under ρ_2 is computed, and the price p_1 or p_2 is offered at time t depending on whether the likelihood ratio is greater or smaller than 1. It was shown in Zhai et al., 2011 [217] and Tehrani, Zhai, and Zhao, 2012 [188] that this strategy offers bounded regret. In other words, with probability one, this dynamic pricing strategy converges to the optimal price within a finite time, thus achieving complete learning. By offering an exploration price—the price that best distinguishes the two demand curves under a distance measure such as the Chernoff distance—when the likelihood ratio is close to 1, regret can be further reduced. It is worth noting that the LRT policy does not require the complete function forms of ρ_1 or ρ_2. It only needs the values of ρ_1 and ρ_2 at the optimal price p_1 of ρ_1 and p_2 of ρ_2. Extensions to an arbitrary but finite number of possible demand curves are straightforward with minor modifications to the policy to preserve a bounded regret.

Bandit formulations of dynamic pricing abound in the literature, within both the Bayesian and the frequentist frameworks and adopting various market models. Representative examples of the former include the work by Aghion, Bolton, and Jullien, 1991 [3], Aviv and Pazgal, 2005 [26], and Harrison, Keskin, and Zeevi, 2011 [99]. In particular, Harrison, Keskin, and Zeevi, while adopting the same Bayesian formulation as in McLennan, 1984 [148], examined the performance of the myopic policy under a regret measure and showed that a modified myopic policy offers bounded regret. Representative studies within the frequentist framework include Kleinberg, 2003 [114] and Besbes and Zeevi, 2009 [37], and the references therein.

6.2.2 WEB SEARCH, ADS DISPLAY, AND RECOMMENDATION SYSTEMS: LEARNING TO RANK

Applications such as Web search, sponsored ads display, and recommendation systems give rise to the problem of learning a list of a few top-ranked items from a large number of alternatives. This leads to the so-called *ranked bandit model*.

Consider the problem of learning the top K most relevant documents for a given query. Let $\mathcal{D}$ denote the set of all documents. Users have different interpretations of the query (consider a query with the keyword "bandit"; users issuing this query may be looking for drastically different things). Each user i is characterized by the set $\mathcal{A}_i \subset \mathcal{D}$ of documents that the user considers relevant. We refer to $\mathcal{A}_i$ as the *type* of a user.

At each time t, a user arrives, with an unknown type drawn i.i.d. from an unknown distribution. The learning policy outputs a ranked list of K documents from $\mathcal{D}$. The user scans the list from the top and clicks the first relevant document, if such documents are presented.

Otherwise, the user abandons the query. This type of user behavior in web search is referred to as the cascade model (see Craswell, et al. [70]). The objective of the online learning policy is to minimize the rate of query abandonment in the long run. This can be modeled by assigning a unit reward for each click and zero reward in the event of abandonment.

In the special case of $K = 1$, this ranked bandit model reduces to the classic bandit model with each document constitutes an independent arm. For $K > 1$, the ranked bandit model can be considered as an instance of the combinatorial bandit model. A unique feature, however, is in the feedback. In the event of a query abandonment, it is known that all K documents in the presented list are irrelevant. When the user clicks the kth document, however, the policy receives no feedback on the relevance of the documents displayed below the kth document. Since the order of the chosen K documents affects only the amount of information in the feedback but not the reward, it might be sufficient to consider an action space consisting of only size-K sets (rather than ordered sets), coupled with an uncertainty measure that allows the policy to display documents with high uncertainty levels (thus more exploration value) at the top.

The complexity of this bandit model can be well appreciated when we set to identify the oracle policy: for a user randomly sampled from a known type distribution, find a set of K documents that maximizes the click-through probability. Note that for the oracle who does not need to learn, ranking is irrelevant.

This problem is a variant of the hitting set problem, a classical NP-complete problem. Specifically, consider a collection of sets $\{\mathcal{A}_i\}$, each a subset of $\mathcal{D}$ and with weight p_i determined by the type distribution. The optimal oracle choice $\mathcal{S}^*$ is a size-K subset of $\mathcal{D}$ that maximizes the total weight of the sets in the collection that are hit by $\mathcal{S}^*$:

$$\mathcal{S}^* = \arg \max_{\mathcal{S} \subset \mathcal{D} : |\mathcal{S}| = K} \sum_i p_i \mathbb{I} \left(\mathcal{S} \bigcap \mathcal{A}_i \neq \emptyset \right). \tag{6.4}$$

The above ranked bandit model was studied by Radlinski, Kleinberg, and Joachims, 2008 [165] and Slivkins, Radlinski, and Gollapudi, 2013 [179]. Streeter and Golovin, 2008 [184] and Streeter, Golovin, and Krause, 2009 [185] considered the problem under general reward functions that are monotone and submodular. Kveton et al., 2015 [123] adopted a more tractable preference model where the user's preference was assumed to be independence across documents/items and over time, and referred to the resulting bandit model as the cascading bandits. Like many other directions for extending the canonical bandit model, the field awaits further exploitation and exploration.

Bibliography

[1] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits, *Proc. of Conference on Neural Information Processing Systems (NeurIPS)*, pages 2312–2320. 96

[2] Agarwal, A., Foster, D. P., Hsu, D., Kakade, S. M., and Rakhlin, A. (2013). Stochastic convex optimization with bandit feedback, *SIAM Journal Optimization*, 23:213–240. DOI: 10.1137/110850827. 98

[3] Aghion, P., Bolton, C. H. P., and Jullien, B. (1991). Optimal learning by experimentation, *The Review of Economic Studies*, 58:621–654. DOI: 10.2307/2297825. 125

[4] Agrawal, R. (1995a). Sample mean based index policies by $O(\log n)$ regret for the multi-armed bandit problem, *Advances in Applied Probability*, 27:1054–1078. DOI: 10.2307/1427934. 5, 65, 73, 75

[5] Agrawal, R. (1995b). The continuum-armed bandit problem, *SIAM Journal on Control and Optimization*, 33:1926–1951. DOI: 10.1137/s0363012992237273. 97

[6] Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem, *Proc. of Conference on Learning Theory (COLT)*. 81, 82, 83

[7] Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for Thompson sampling, *Proc. of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 81, 83

[8] Agrawal, R. and Teneketzis, D. (1989). Certainty equivalence control with forcing: Revisited. *Systems and Control Letters*, 13(5):405–412, December 1989. DOI: 10.1016/0167-6911(89)90107-2. 74

[9] Ahmad, S. H., Liu, M., Javadi, T., Zhao, Q., and Krishnamachari, B. (2009). Optimality of myopic sensing in multi-channel opportunistic access, *IEEE Transactions on Information Theory*, 55:4040–4050. DOI: 10.1109/tit.2009.2025561. 119

[10] Albert, A. E. (1961). The sequential design of experiments for infinitely many states of nature, *The Annals of Mathematics Statistics*, 32:774–799. DOI: 10.1214/aoms/1177704973. 111

[11] Allenberg, C., Auer, P., Gyorfi, L., and Ottucsák, G. (2006). Hannan consistency in on-line learning in case of unbounded losses under partial monitoring, *Proc. of International Conference on Algorithmic Learning Theory (ALT)*, pages 229–243. DOI: 10.1007/11894841_20. 101

[12] Allesiardo, R. and Feraud, R. (2015). Exp3 with drift detection for the switching bandit problem, *Proc. of IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. DOI: 10.1109/dsaa.2015.7344834. 91

[13] Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits, *Proc. of the 28th Conference on Learning Theory (COLT)*, 40:23–35. 101

[14] Anandkumar, A., Michael, N., Tang, A. K., and Swami, A. (2011). Distributed algorithms for learning and cognitive medium access with logarithmic regret, *IEEE Journal on Selected Areas in Communications*, 29:731–745. DOI: 10.1109/jsac.2011.110406. 116

[15] Anantharam, V., Varaiya, P., and Walrand, J. (1987a). Asymptotically efficient allocation rules for the multi-armed bandit problem with multiple plays—Part I: I.I.D. rewards, *IEEE Transactions on Automatic Control*, 32:968–975. DOI: 10.1109/TAC.1987.1104491. 88, 115

[16] Anantharam, V., Varaiya, P., and Walrand, J. (1987b). Asymptotically efficient allocation rules for the multi-armed bandit problem with multiple plays-Part II: Markovian rewards, *IEEE Transaction on Automatic Control*, 32:977–982. DOI: 10.1109/tac.1987.1104485. 87

[17] Arrow, K. (1951). *Social Choice and Individual Values*, John Wiley & Sons. 102

[18] Asawa, M. and Teneketzis, D. (1996). Multi-armed bandits with switching penalties, *IEEE Transactions on Automatic Control*, 41:328–348. DOI: 10.1109/9.486316. 52

[19] Audibert, J. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits, *Proc. of the 22nd Annual Conference on Learning Theory (COLT)*. 79, 93

[20] Audibert, J. and Bubeck, S. (2010). Regret bounds and minimax policies under partial monitoring, *Journal of Machine Learning Research*, pages 2785–2836. 101

[21] Audibert, J., Munos, R., and Szepesvári, C. (2009). Exploration-exploitation trade-off using variance estimates in multi-armed bandits, *Theoretical Computer Science*, pages 1876–1902. DOI: 10.1016/j.tcs.2009.01.016. 108

[22] Audibert, J., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits, *Proc. of the 23rd Annual Conference on Learning Theory (COLT)*. 111

[23] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem, *Machine Learning*, 47:235–256. DOI: 10.1023/A:1013689704352. 5, 74, 75, 76

[24] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R.E. (2003). The non-stochastic multi-armed bandit problem, *SIAM Journal on Computing*, 32:48–77. DOI: 10.1137/s0097539701398375. 88, 90, 93

[25] Auer, P., Ortner, R., and Szepesvári, C. (2007). Improved rates for the stochastic continuum-armed bandit problem, *Proc. of the 23rd Annual Conference on Learning Theory (COLT)*, pages 454–468. DOI: 10.1007/978-3-540-72927-3_33. 98

[26] Aviv, Y. and Pazgal, A. (2005). A partially observed Markov decision process for dynamic pricing, *Management Science*, 51:1400–1416. DOI: 10.1287/mnsc.1050.0393. 125

[27] Awerbuch, B. and Kleinberg, R. (2008). Online linear optimization and adaptive routing, *Journal of Computer and System Sciences*, pages 97–114. DOI: 10.1016/j.jcss.2007.04.016. 121

[28] Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2013). Bandits with knapsacks, *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 207–216. DOI: 10.1109/focs.2013.30. 99

[29] Banks, J. and Sundaram, R. (1994). Switching costs and the Gittins index, *Econometrica*, 62:687–694. DOI: 10.2307/2951664. 51

[30] Bellman, R. (1956). A problem in the sequential design of experiments, *Sankhia*, 16:221–229. 1, 3, 18, 109

[31] Bellman, R. (1957). *Dynamic Programming*, Princeton University Press. 54

[32] Ben-Israel, A. and Flåm, S. D. (1990). A bisection/successive approximation method for computing Gittins indices, *Methods and Models of Operations Research*, 34:411. DOI: 10.1007/bf01421548. 24

[33] Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*, Springer. DOI: 10.1007/978-94-015-3711-7. 2, 6, 34

[34] Bertsekas, D. P. (1992). Auction algorithms for network flow problems: A tutorial introduction, *Computational Optimization Applications*, 1:7–66. DOI: 10.1007/bf00247653. 116

[35] Bertsimas, D. and Nino-Mora, J. (2000). Restless bandits, linear programming relaxations and a primal-dual index heuristic, *Operations Research*, 48:80–90. DOI: 10.1287/opre.48.1.80.12444. 50

[36] Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards, *Proc. of the 27th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 199–207. 91, 92

[37] Besbes, O. and Zeevi, A. (2009). Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms, *Operations Research*, 57:1407–1420. DOI: 10.1287/opre.1080.0640. 125

[38] Bessler, S. (1960). Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments: Part I–Theory, *Technical Representative Applied Mathematics and Statistics Laboratories*, Stanford University. 111

[39] Bistritz, I. and Leshem, A. (2018). Distributed multi-player bandits—a game of thrones approach, *Proc. of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. 116

[40] Blackwell, D. (1956). An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8. DOI: 10.2140/pjm.1956.6.1. 100

[41] Blackwell, D. (1962). Discrete dynamic programming, *Annals of Mathematical Statistics*, 32:719–726. DOI: 10.1214/aoms/1177704593. 55

[42] Bradfield, J. (2007). *Introduction to the Economics of Financial Markets*, Oxford University Press. 105

[43] Bradt, R. N., Johnson, S. M., and Karlin, S. (1956). On sequential designs for maximizing the sum of *n* observations, *Annals of Mathematical Statistics*, 27:1060–1074. DOI: 10.1214/aoms/1177728073. 1, 18, 109

[44] Brown, D. B. and Smith, J. E. (2013). Optimal sequential exploration: Bandits, clairvoyants, and wildcats. *Operations Research*, 61(3):644–665. 38

[45] Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems, *Foundations and Trends in Machine Learning*. DOI: 10.1561/2200000024. 66, 79

[46] Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail, *IEEE Transactions on Information Theory*, 59:7711–7717. DOI: 10.1109/tit.2013.2277869.

[47] Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure exploration in finitely-armed and continuous-armed bandits, *Theoretical Computer Science*, 412:1832–1852. DOI: 10.1016/j.tcs.2010.12.059. 112

[48] Bubeck, S., Munos, R., Stoltz, G., and Szepesvari, C. (2011). Online optimization in X-armed bandits, *Journal of Machine Learning Research (JMLR)*, 12:1587–1627. 98

[49] Bubeck, S., Stoltz, G., and Yu, J. Y. (2011). Lipschitz bandits without the Lipschitz constant, *Proc. of the 22nd International Conference on Algorithmic Learning Theory (ALT)*, pages 144–158. DOI: 10.1007/978-3-642-24412-4_14. 98

[50] Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail, *IEEE Transactions on Information Theory*, 59:7711–7717. DOI: 10.1109/tit.2013.2277869. 79

[51] Bubeck, S., Li, Y., Peres, Y., and Sellke, M. (2019). Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. *ArXiv Preprint*, arXiv:1904.12233. 116

[52] Buccapatnam, S., Eryilmaz, A., and Shroff, N. B. (2014). Stochastic bandits with side observations on networks. *ACM SIGMETRICS Performance Evaluation Review*, 42(1):289–300. 100

[53] Bull, A. (2015). Adaptive-treed bandits, *Bernoulli Journal of Statistics*, 21:2289–2307. DOI: 10.3150/14-bej644. 98

[54] Cappé, O., Garivier, A., Maillard, O., Munos, R., and Stoltz, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation, *The Annals of Statistics*, 41(3):1516–1541. DOI: 10.1214/13-aos1119. 73

[55] Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. (2012). Leveraging side observations in stochastic bandits, *Proc. of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 142–151. 100

[56] Cesa-Bianchi, N., Lugosi, G., and Stoltz, G. (2005). Minimizing regret with label-efficient prediction, *IEEE Transactions on Information Theory*, 51:2152–2162. DOI: 10.1109/tit.2005.847729. 101

[57] Cesa-Bianchi, N., Gentile, C., and Zappella, G. (2013). A gang of bandits, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pages 737–745.

[58] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*, Cambridge University Press. DOI: 10.1017/cbo9780511546921. 100

[59] Chakravorty, J. and Mahajan, A. (2014). Multi-armed bandits, Gittins index, and its calculation, *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods*, N. Balakrishnan, Ed., John Wiley & Sons, Inc. DOI: 10.1002/9781118596333.ch24. 24

[60] Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications, *Proc. of the 30th International Conference on Machine Learning (ICML)*, pages 151–159. 95, 96

[61] Chen, Y. and Katehakis, M. (1986). Linear programming for finite state multi-armed bandit problems, *Mathematics of Operations Research*, 11:180–183. DOI: 10.1287/moor.11.1.180. 26

[62] Chernoff, H. (1959). Sequential design of experiments, *The Annals of Mathematical Statistics*, 30:755–770. DOI: 10.1214/aoms/1177706205. 109, 111

[63] Chow, Y. and Lai, T. (1975). Some one-sided theorems on the tail distribution of sample sums with applications to the last time and largest excess of boundary crossings, *Transactions of the American Mathematical Society*, 208:51–72. DOI: 10.2307/1997275. 71

[64] Chu, W., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandits with linear payoff functions, *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*. 114

[65] Cohen, A., Hazan, T., and Koren, T. (2016). Online learning with feedback graphs without the graphs, *Proc. of The 33rd International Conference on Machine Learning (ICML)*. 101

[66] Combes, R. and Proutiere, A. (2014). Unimodal bandits: Regret lower bounds and optimal algorithms. *Proc. of the International Conference on Machine Learning (ICML)*, pages 521–529. 98

[67] Combes, R., Jiang, C., and Srikant, R. (2015). Bandits with budgets: Regret lower bounds and optimal algorithms, *Proc. of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 245–257. DOI: 10.1145/2745844.2745847. 99

[68] Combes, R., Magureanu, S., and Proutiere, A. (2017). Minimal exploration in structured stochastic bandits, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pages 1763–1771. 99

[69] Cope, E. W. (2009). Regret and convergence bounds for a class of continuum-armed bandit problems, *IEEE Transactions on Automatic Control*, 54. DOI: 10.1109/tac.2009.2019797. 98

[70] Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models, *Proc. of the 1st ACM International Conference on Web Search and Data Mining*, pages 87–94. 126

[71] Dai, W., Gai, Y., Krishnamachari, B., and Zhao, Q. (2011). The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2940–2943. DOI: 10.1109/icassp.2011.5946273. 120

[72] Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback, *Proc. of the Annual Conference on Learning Theory (COLT)*, pages 355–366. 96, 98

[73] Denardo, E. V., Park, H., and Rothblum, U. G., (2007). Risk-sensitive and risk-neutral multiarmed bandits, *Mathematics of Operations Research*. DOI: 10.1287/moor.1060.0240. 26

[74] Denardo, E. V., Feinberg, E. A., and Rothblum, U. G. (2013). The multi-armed bandit, with constraints, *Annals of Operations Research*, 1:37–62. DOI: 10.1007/s10479-012-1250-y. 26

[75] Dumitriu, I., Tetali, P., and Winkler, P. (2003). On playing golf with two balls, *SIAM Journal of Discrete Mathematics*, 4:604–615. DOI: 10.1137/s0895480102408341. 52, 53

[76] Ehsan, N. and Liu, M. (2004). On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services, *IEEE International Conference on Computer Communications (INFOCOM)*. DOI: 10.1109/infcom.2004.1354606. 50

[77] Even-Dar, E., Kearns, M., and Wortman, J. (2006). Risk-sensitive online learning, *Proc. of the 17th International Conference on Algorithmic Learning Theory (ALT)*, pages 199–213. DOI: 10.1007/11894841_18. 108

[78] Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems, *Journal of Machine Learning Research*, 7:1079–1105. 111

[79] Feldman, D. (1962). Contributions to the two-armed bandit problem, *Annals of Mathematical Statistics*, 3:847–856. DOI: 10.1214/aoms/1177704454. 1

[80] French, K. R., Shwert, G. W., and Stambaugh, R. F. (1987). Expected stock returns and volatility, *Journal of Financial Economics*, 19:3–29. DOI: 10.1016/0304-405x(87)90026-2. 105

[81] Gai, Y. and Krishnamachari, B. (2014). Distributed stochastic online learning policies for opportunistic spectrum access, *IEEE Transactions on Signal Processing*. DOI: 10.1109/tsp.2014.2360821. 116

[82] Gai, Y., Krishnamachari, B., and Jain, R. (2011). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations, *IEEE/ACM Transactions on Networking*, 5:1466–1478. DOI: 10.1109/tnet.2011.2181864. 95, 96, 116, 121

[83] Galichet, N., Sebag, M., and Teytaud, O. (2013). Exploration vs. exploitation vs. safety: Risk-averse multi-armed bandits, *Proc. of the Asian Conference on Machine Learning*. 108

[84] Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond, *Proc. of the Annual Conference on Learning Theory (COLT)*. 73

[85] Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems, *Proc. of the International Conference on Algorithmic Learning Theory (ALT)*, pages 174–188. DOI: 10.1007/978-3-642-24412-4_16. 90

[86] Gilbert, E. N. (1960). Capacity of burst-noise channel, *Bell System Technical Journal*, 39:1253–1265. DOI: 10.1002/j.1538-7305.1960.tb03959.x. 117

[87] Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments, *Progress in Statistics*, pages 241–266, read at the 1972 European Meeting of Statisticians, Budapest. 3, 11, 13, 22, 56

[88] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices, *Journal of the Royal Statistical Society*, 2:148–177. DOI: 10.1111/j.2517-6161.1979.tb01068.x. 3, 13, 22

[89] Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices*, John Wiley & Sons, Inc. DOI: 10.1002/9780470980033. 2

[90] Gittins, J. C., Glazebrook, K. D., and Weber, R. R. (2011). *Multi-Armed Bandit Allocation Indices*, John Wiley & Sons, Inc. DOI: 10.1002/9780470980033. 4, 13, 15, 22, 29, 34, 40

[91] Glazebrook, K. D. (1976). Stochastic scheduling with order constraints, *International Journal of Systems Science*, 6:657–666. DOI: 10.1080/00207727608941950. 41

[92] Glazebrook, K. D. (1979). Stoppable families of alternative bandit processes, *Journal of Applied Probability*, 16(4):843–854. DOI: 10.2307/3213150. 38

[93] Glazebrook, K. D. (1982). On a sufficient condition for superprocesses due to Whittle. *Journal of Applied Probability*, pages 99–110. 38

[94] Glazebrook, K. D., Ruiz-Hernandez, D., and Kirkbride, C. (2006). Some indexable families of restless bandit problems, *Advances in Applied Probability*, pages 643–672. DOI: 10.1239/aap/1158684996. 46, 50

[95] Glazebrook, K. D. and Ruiz-Hernández, D. (2005). A restless bandit approach to stochastic scheduling problems with switching costs. `https://www.researchgate.net/publication/252641987_A_Restless_Bandit_Approach_to_Stochastic_Scheduling_Problems_with_Switching_Costs` 52

[96] Hadfield-Menell, D. and Russell, S. (2015). Multitasking: Efficient optimal planning for bandit superprocesses. *Proc. of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 345–354. 38

[97] Hanawal, M. K. and Saligrama, V. (2015). Efficient detection and localization on graph structured data, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5590–5594. DOI: 10.1109/icassp.2015.7179041. 96, 97

[98] Hannan, J. (1957). Approximation to Rayes risk in repeated play, *Contributions to the Theory of Games*, 3:97–139. DOI: 10.1515/9781400882151-006. 61, 100

[99] Harrison, J. M., Keskin, N. B., and Zeevi, A. (2011). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution, *Management Science*, pages 1–17. DOI: 10.1287/mnsc.1110.1426. 125

[100] Hartland, C., Baskiotis, N., Gelly, S., Sebag, M., and Teytaud, O. (2007). Change point detection and meta-bandits for online learning in dynamic environments, CAp, cepadues, pages 237–250. 91

[101] Helmbold, D. and Panizza, S. (1997). Some label efficient learning results, *Proc. of the Annual Conference on Learning Theory (COLT)*. DOI: 10.1145/267460.267502. 101

[102] Honda, J. and Takemura, A. (2014). Optimality of Thompson sampling for Gaussian bandits depends on priors. *Proc. of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 83

[103] Jiang, C. and Srikant, R. (2003). Bandits with budgets, *Proc. of the IEEE Conference on Decision and Control (CDC)*, pages 5345–5350. DOI: 10.1109/cdc.2013.6760730. 99

[104] Jun, T. (2004). A survey on the bandit problem with switching costs, *De Economist*, 4:513–541. DOI: 10.1007/s10645-004-2477-z. 52

[105] Kalathil, D., Nayyar, N., and Jain, R. (2014). Decentralized learning for multi-player multi-armed bandits, *IEEE Transactions on Information Theory*. DOI: 10.1109/cdc.2012.6426587. 116

[106] Kallenberg, L. C. M. (1986). A note on M. N. Katehakis' and Y.-R. Chen's computation of the Gittins index. *Mathematics of Operations Research*, 11(1):184–186. DOI: 10.1287/moor.11.1.184. 26

[107] Kanade, V., McMahan, B., and Bryan, B. (2009). Sleeping experts and bandits with stochastic action availability and adversarial rewards, *Proc. of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 99

[108] Kaufmann, E. (2018). On Bayesian index policies for sequential resource allocation, *The Annals of Statistics*, 46(2):842–865. DOI: 10.1214/17-aos1569. 83

[109] Kaufmann, E., Cappe, O., and Garivier, A. (2012). On Bayesian upper confidence bounds for bandit problems, *Proc. of the 15th International Conference on Artificial Intelligence and Statistics (AISTAT)*. 83

[110] Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An optimal finite time analysis, *Proc. of International Conference on Algorithmic Learning Theory (ALT)*. 83

[111] Katehakis, M. N. and Rothblum, U. G. (1996). Finite state multi-armed bandit problems: Sensitive-discount, average-reward and average-overtaking optimality, *The Annals of Applied Probability*, 6(3):1024–1034. DOI: 10.1214/aoap/1034968239. 55

[112] Katehakis, M. and Veinott, A. (1987). The multi-armed bandit problem: Decomposition and computation, *Mathematics of Operations Research*, 12(2):262–268. DOI: 10.1287/moor.12.2.262. 19

[113] Kelly, F. P. (1979). Discussion of Gittins' paper, *Journal of the Royal Statistical Society*, 41(2):148–177. 54

[114] Kleinberg, R. (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions, *Proc. of the 44th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 594–605. DOI: 10.1109/sfcs.2003.1238232. 125

[115] Kleinberg, R. (2004). Nearly tight bounds for the continuum-armed bandit problem, *Proc. of the 17th International Conference on Neural Information Processing Systems*, pages 697–704. 97

[116] Kleinberg, R., Niculescu-Mizil, A., and Sharma, Y. (2008). Regret bounds for sleeping experts and bandits, *Proc. of the Annual Conference on Learning Theory (COLT)*. DOI: 10.1007/s10994-010-5178-7. 99

[117] Kleinberg, R., Slivkins, A., and Upfal, E. (2015). Bandits and experts in metric spaces. `https://arxiv.org/abs/1312.1277` DOI: 10.1145/3299873. 98, 114

[118] Klimov, G. P. (1974). Time-sharing service systems, *Teor. Veroyatnost. i Primenen.*, 19(3):558–576, *Theory of Probability and its Applications*, 19(3):532–551, 1975. DOI: 10.1137/1119060 .

[119] Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. (2015). Regret lower bound and optimal algorithm in dueling bandit problem, *Proc. of Conference on Learning Theory (COLT)*. 103

[120] Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*. 83

[121] Krishnamurthy, V. and Wahlberg, B. (2009). Partially observed Markov decision process multiarmed bandits—structural results, *Mathematics of Operations Research*, 34(2):287–302. DOI: 10.1287/moor.1080.0371. 24

[122] Kuhn, H. W. (1955). The Hungarian method for the assignment problem, *Naval Res. Logistics Quart.*, 2(1–2):83–97. 116

[123] Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. (2015). Cascading bandits: learning to rank in the cascade model, *Proc. of the 32nd International Conference on Machine Learning (ICML)*. 126

[124] Kveton, B., Wen, Z., Ashkan, A., Ashkan, A., and Szepesvári, C. (2015). Tight regret bounds for stochastic combinatorial semi-bandits, *Proc. of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 535–543. 95, 96

[125] Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem, *The Annals of Statistics*, 15(3):1091–1114. DOI: 10.1214/aos/1176350495. 73, 80

[126] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, 6(1):4–22. DOI: 10.1016/0196-8858(85)90002-8. 5, 71, 72, 87, 88, 106, 115

[127] Lai, T. L. and Ying, Z. (1988). Open bandit processes and optimal scheduling of queueing networks, *Advances in Applied Probability*, 20(2):447–472. DOI: 10.2307/1427399. 42

[128] Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for contextual multi-armed bandits, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*. 114

[129] Lattimore, T. and Szepesvári, C. (2019). *Bandit Algorithms*, To be published by Cambridge University Press. `https://tor-lattimore.com/downloads/book/book.pdf` 79, 93

[130] Le Ny, J., Dahleh, M., and Feron, E. (2008). Multi-UAV dynamic routing with partial observations using restless bandit allocation indices, *Proc. of the American Control Conference*. DOI: 10.1109/acc.2008.4587156. 117, 118

[131] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation, *WWW*, pages 661–670. DOI: 10.1145/1772690.1772758. 114

[132] Liu, H., Liu, K., and Zhao, Q. (2013). Learning in a changing world: Restless multiarmed bandit with unknown dynamics, *IEEE Transactions on Information Theory*, 59(3):1902–1916. DOI: 10.1109/tit.2012.2230215. 89

[133] Liu, F., Lee, J., and Shroff, N. B. (2018). A change-detection based framework for piecewise-stationary multi-armed bandit problem, *Proc. of the AAAI Conference on Artificial Intelligence*. 91

[134] Liu, K. and Zhao, Q. (2008). A restless bandit formulation of opportunistic access: Indexablity and index policy, *Proc. of the 5th IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON) Workshops*. DOI: 10.1109/sahcnw.2008.12. 118

[135] Liu, K. and Zhao, Q. (2010). Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access, *IEEE Transactions on Information Theory*, 56(11):5547–5567. 118, 119

[136] Liu, K. and Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players, *IEEE Transactions on Signal Processing*, 58(11):5667–5681. DOI: 10.1109/tsp.2010.2062509. 116

[137] Liu, K. and Zhao, Q. (2012). Adaptive shortest-path routing under unknown and stochastically varying link states, *IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pages 232–237. 95, 96, 121

[138] Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem, *Proc. of the 33rd International Conference on Machine Learning (ICML)*, 48:1690–1698. 112

[139] Lott, C. and Teneketzis, D. (2000). On the optimality of an index rule in multi-channel allocation for single-hop mobile networks with multiple service classes, *Probability in the Engineering and Informational Sciences*, 14:259–297. DOI: 10.1017/s0269964800143013. 50

[140] Maillard, O., Munos, R., and Stoltz, G. (2011). Finite-time analysis of multi-armed bandits problems with Kullback–Leibler divergences, *Proc. of Conference On Learning Theory (COLT)*. 73

[141] Maillard, O. (2013). Robust risk-averse stochastic multi-armed bandits, *Proc. of the International Conference on Algorithmic Learning Theory (ALT)*, 8139:218–233. DOI: 10.1007/978-3-642-40935-6_16. 108

[142] Mandelbaum, A. (1986). Discrete multi-armed bandits and multi-parameter processes, *Probability Theory and Related Fields*, 71(1):129–147. DOI: 10.1007/bf00366276. 20, 21

[143] Mandelbaum, A. (1987). Continuous multi-armed bandits and multiparameter processes, *The Annals of Probability*, 15(4):1527–1556. DOI: 10.1214/aop/1176991992.

[144] Mannor, S. and Shamir, O. (2011). From bandits to experts: On the value of side-observations, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 24:684–692. 100

[145] Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem, *Journal of Machine Learning Research*, 5:623–648. 111

[146] Marden, J. R., Young, H. P., and Pao L. Y. (2014). Achieving pareto optimality through distributed learning, *SIAM Journal on Control and Optimization*, 52(5):2753–2770. DOI: 10.1137/110850694. 116

[147] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91. 103

[148] McLennan, A. (1984). Price dispersion and incomplete learning in the long run, *Journal of Economic Dynamics and Control*, 7(3):331–347. DOI: 10.1016/0165-1889(84)90023-x. 124, 125

[149] Mellor, J. and Shapiro, J. (2013). Thompson sampling in switching environments with Bayesian online change detection, *Proc. of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 442–450. 91

[150] Minsker, S. (2013). Estimation of extreme values and associated level sets of a regression function via selective sampling, *Proc. of the 26th Conference on Learning Theory (COLT)*, pages 105–121. 98

[151] Naghshvar, M. and Javidi, T. (2013). Active sequential hypothesis testing, *Annals of Statistics*, 41(6). DOI: 10.1214/13-aos1144. 111

[152] Nain, P., Tsoucas, P., and Walrand, J. (1989). Interchange arguments in stochastic scheduling, *Journal of Applied Probability*, 26(4):815–826. DOI: 10.21236/ada454729.

[153] Nash, P. (1973). Optimal allocation of resources between research projects, Ph.D. thesis, Cambridge University. 42

[154] Niño-Mora, J. (2001). Restless bandits, partial conservation laws and indexability, *Advances in Applied Probability*, 33:76–98. DOI: 10.1017/S0001867800010648. 46, 50

[155] Niño-Mora, J. (2002). Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach, *Mathematical Programming*, 93:361–413. DOI: 10.1007/s10107-002-0362-6.

[156] Niño-Mora, J. (2006). Restless bandit marginal productivity indices, diminishing returns and optimal control of make-to-order/make-to-stock M/G/1 queues, *Mathematics Operations Research*, 31:50–84. DOI: 10.1287/moor.1050.0165. 50

[157] Niño-Mora, J. (2007a). A $(2/3)n^3$ fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain, *INFORMS Journal on Computing*, 19(4):485–664. DOI: 10.1287/ijoc.1060.0206. 26

[158] Niño-Mora, J. (2007b). Dynamic priority allocation via restless bandit marginal productivity indices, *TOP*, 15:161–198. DOI: 10.1007/s11750-007-0025-0. 46, 50

[159] Niño-Mora, J. (2008). A faster index algorithm and a computational study for bandits with switching costs, *INFORMS Journal of Computation*, 20:255–269. DOI: 10.1287/ijoc.1070.0238. 52

[160] Niño-Mora, J. (2011). Computing a classic index for finite-horizon bandits, *INFORMS Journal of Computation*, 23:254–267. DOI: 10.1287/ijoc.1100.0398. 56

[161] Nitinawarat, S., Atia, G. K., and Veeravalli, V. V. (2013). Controlled sensing for multihypothesis testing, *IEEE Transactions on Automatic Control*, 58(10):2451–2464. DOI: 10.1109/tac.2013.2261188. 111

[162] Papadimitriou, C. H. and Tsitsiklis, J. N. (1999). The complexity of optimal queueing network control, *Mathematics of Operations Research*, 24(2):293–305. DOI: 10.1109/sct.1994.315792. 88

[163] Pradelski, B. S. and Young, H. P. (2012). Learning efficient Nash equilibria in distributed systems, *Games and Economic Behavior*, 75(2):882–897. DOI: 10.1016/j.geb.2012.02.017. 116

[164] Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley. DOI: 10.1002/9780470316887. 7, 8, 10, 27, 52

[165] Radlinski, F., Kleinberg, R., and Joachims, T. (2008). Learning diverse rankings with multiarmed bandits, *Proc. of International Conference on Machine Learning (ICML)*, pages 784–791. DOI: 10.1145/1390156.1390255. 126

[166] Raghunathan, V., Borkar, V., Cao, M., and Kumar, P. R. (2008). Index policies for real-time multicast scheduling for wireless broadcast systems, *Proc. of IEEE International Conference on Computer Communications (INFOCOM)*. DOI: 10.1109/infocom.2007.217. 50

[167] Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates, *Proc. of the Annual Conference on Learning Theory (COLT)*. 114

[168] Robbins, H. (1952). Some aspects of the sequential design of experiments, *Bulletin of the American Mathematical Society*, 58(5):527–535. DOI: 10.1090/s0002-9904-1952-09620-8. 1, 5, 74, 109

[169] Rosenski, J., Shamir, O., and Szlak, L. (2016). Multi-player bandits—a musical chairs approach, *Proc. of International Conference on Machine Learning (ICML)*, pages 155–163. 116

[170] Ross, S. M. (1970). *Applied Probability Models with Optimization Applications*, Holden Day. 27

[171] Ross, S. M. (1995). *Introduction to Stochastic Dynamic Programming*, Academic Press. DOI: 10.1016/C2013-0-11415-8. 7

[172] Rothschild, M. (1974). A two-armed bandit theory of market pricing, *Journal of Economic Theory*, 9(2):185–202. DOI: 10.1016/0022-0531(74)90066-0. 123

[173] Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits, *Mathematics of Operations Research*, 35(2):395–411. DOI: 10.1287/moor.1100.0446. 96

[174] Salomon, A. and Audibert, J. Y. (2011). Deviations of stochastic bandit regret, *Proc. of the International Conference on Algorithmic Learning Theory (ALT)*. DOI: 10.1007/978-3-642-24412-4_15. 108

[175] Sani, A., Lazaric, A., and Munos, R. (2012). Risk aversion in multi-armed bandits, *Proc. of Neural Information Processing Systems (NeurIPS)*. 105, 107

[176] Seldin, Y., Auer, P., Laviolette, F., Shawe-Taylor, J., and Ortner, R. (2011). PAC-Bayesian analysis of contextual bandits, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*. 114

[177] Slivkins, A. (2011). Multi-armed bandits on implicit metric spaces, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*. 98

[178] Slivkins, A. (2014). Contextual bandits with similarity information, *Journal of Machine Learning Research*, 15:2533–2568. 114

[179] Slivkins, A., Radlinski, F., and Gollapudi, S. (2013). Ranked bandits in metric spaces: Learning diverse rankings over large document collections, *Journal of Machine Learning Research*. 126

[180] Slivkins, A. and Upfal, E. (2008). Adapting to a changing environment: The Brownian restless bandits, *Proc. of the Annual Conference on Learning Theory (COLT)*. 91

[181] Smallwood, R. and Sondik, E. (1971). The optimal control of partially ovservable Markov processes over a finite horizon, *Operations Research*, pages 1071–1088. DOI: 10.1287/opre.21.5.1071. 118

[182] Sonin, I. M. (2008). A generalized Gittins index for a Markov chain and its recursive calculation, *Statistics and Probability Letters*, 78:1526–1533. DOI: 10.1016/j.spl.2008.01.049. 26

[183] Stoltz, G. (2005). Incomplete information and internal regret in prediction of individual sequences. Ph.D. thesis, Université Paris Sud-Paris XI. 93

[184] Streeter, M. and Golovin, D. (2008). An online algorithm for maximizing submodular functions, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*. DOI: 10.21236/ada476748. 126

[185] Streeter, M., Golovin, D., and Krause, A. (2009). Online learning of assignments, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*. 126

[186] Sundaram, R. K. (2005). Generalized bandit problems, *Social Choice and Strategic Decisions: Studies in Choice and Welfare*, Austen-Smith, D. and Duggan, J., Eds., Springer, Berlin, Heidelberg. DOI: 10.1007/3-540-27295-x_6. 52

[187] Sutton, R. and Barto, A. (1998). *Reinforcement Learning, an Introduction*. Cambridge, MIT Press/Bradford Books. 74

[188] Tehrani, P., Zhai, Y., and Zhao, Q. (2012). Dynamic pricing under finite space demand uncertainty: A multi-armed bandit with dependent arms. `https://arxiv.org/abs/1206.5345` 124, 125

[189] Tekin, C. and Liu, M. (2012). Online learning of rested and restless bandits, *IEEE Transaction on Information Theory*, 58(8):5588–5611. DOI: 10.1109/tit.2012.2198613. 88, 89

[190] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika*, 25(3/4):275–294. DOI: 10.1093/biomet/25.3-4.285. 1, 12, 81, 109, 123

[191] Tran-Thanh, L., Chapman, A., de Cote, E. M., Rogers, A., and Jennings, N. R. (2010). Epsilon-first policies for budget-limited multi-armed bandits, *Proc. of the AAAI Conference on Artificial Intelligence*. 99

[192] Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. R. (2012). Knapsack based optimal policies for budget-limited multi-armed bandits, *Proc. of the AAAI Conference on Artificial Intelligence*. 99

[193] Vakili, S., Boukouvalas, A., and Zhao, Q. (2019). Decision variance in risk-averse online learning, *Proc. of the IEEE Conference on Decision and Control (CDC)*. 106

[194] Vakili, S., Liu, K., and Zhao, Q. (2013). Deterministic sequencing of exploration and exploitation for multi-armed bandit problems, *IEEE Journal of Selected Topics in Signal Processing*, 7(5):759–767. DOI: 10.1109/jstsp.2013.2263494. 74, 75, 79, 116

[195] Vakili, S. and Zhao, Q. (2015). Mean-variance and value at risk in multi-armed bandit problems, *Proc. of the 53rd Annual Allerton Conference on Communication, Control, and Computing*. DOI: 10.1109/allerton.2015.7447162. 106, 108

[196] Vakili, S. and Zhao, Q. (2016). Risk-averse multi-armed bandit problems under mean-variance measure, *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111. DOI: 10.1109/jstsp.2016.2592622. 105, 106, 107

[197] Vakili, S., Zhao, Q., Liu, C., and Chuah, C. N. (2018). Hierarchical heavy hitter detection under unknown models, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 122, 123

[198] Valko, M., Carpentier, A., and Munos, R. (2013). Stochastic simultaneous optimistic optimization, *Proc. of the 30th International Conference on Machine Learning (ICML)*, pages 19–27. 98

[199] Valko, M., Munos, R., Kveton, B., and Kocák, T. (2014). Spectral bandits for smooth graph functions, *Proc. of the 31th International Conference on Machine Learning (ICML)*. 96, 97

[200] Varaiya, P., Walrand, J., and Buyukkoc, C. (1985). Extensions of the multiarmed bandit problem: The discounted case, *IEEE Transactions on Automatic Control*, 30(5):426–439. DOI: 10.1109/tac.1985.1103989. 21, 24

[201] Veatch, M. L. and Wein, M. (1996). Scheduling a make-to-stock queue: Index policies and hedging points, *Operations Research*, 44:634–647. DOI: 10.1287/opre.44.4.634. 50

[202] Veinott, A. F., JR. (1966). On finding optimal policies in discrete dynamic programming with no discounting, *Annals of Mathematical Statistics*, 37:1284–1294. DOI: 10.1214/aoms/1177699272. 55

[203] Vogel, W. (1960a). A sequential design for the two armed bandit, *The Annals of Mathematical Statistics*, 31(2):430–443. DOI: 10.1214/aoms/1177705906. 1

[204] Vogel, W. (1960b). An asymptotic minimax theorem for the two armed bandit problem, *Annals of Mathematical Statistics*, 31(2):444–451. DOI: 10.1214/aoms/1177705907. 1, 6, 66

[205] Wang, C. C., Kulkarni, S. R., and Poor, H. V. (2005). Bandit problems with side observations, *IEEE Transactions on Automatic Control*, 50(3):338–355. DOI: 10.1109/tac.2005.844079. 113

[206] Weber, R. R. and Weiss, G. (1990). On an index policy for restless bandits, *Journal of Applied Probability*, 27(3):637–648. DOI: 10.2307/3214547. 4, 50

[207] Weber, R. R. and Weiss, G. (1991). Addendum to on an index policy for restless bandits, *Advances in Applied Probability*, 23(2):429–430. DOI: 10.2307/1427757. 50

[208] Weber, R. R. (1992). On the Gittins index for multiarmed bandits, *The Annals of Applied Probability*, 2(4):1024–1033. DOI: 10.1214/aoap/1177005588. 4, 18, 22

[209] Weiss, G. (1988). Branching bandit processes, *Probability in the Engineering and Informational Sciences*, 2(3):269–278. DOI: 10.1017/s0269964800000826. 42

[210] Whittle, P. (1980). Multi-armed bandits and the Gittins index, *Journal of the Royal Statistical Society*, series B, 42(2):143–149. DOI: 10.1111/j.2517-6161.1980.tb01111.x. 19, 37, 38

[211] Whittle, P. (1981). Arm-acquiring bandits, *The Annals of Probability*, 9(2):284–292. DOI: 10.1214/aop/1176994469. 42

[212] Whittle, P. (1988). Restless bandits: Activity allocation in a changing world, *Journal of Applied Probability*, 25:287–298. DOI: 10.1017/s0021900200040420. 4, 34, 42, 46

[213] Wu, H. and Liu, X. (2016). Double Thompson sampling for dueling bandits, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*. 103

[214] Xu, X., Vakili, S., Zhao, Q., and Swami, A. (2019). Multi-armed bandits on partially revealed unit interval graphs, *IEEE Transactions on Network Science and Engineering*. DOI: 10.1109/TNSE.2019.2935256. 98

[215] Yakowitz, S. and Lowe, W. (1991). Nonparametric bandit methods. *Annals of Operations Research*, 28(1–4):297–312. DOI: 10.1007/BF02055587. 74

[216] Yue, Y. Broder, J., Kleinberg, R., and Joachims, T. (2012). The K-armed dueling bandits problem, *Journal of Computer and System Sciences*, 78:1538–1556. DOI: 10.1016/j.jcss.2011.12.028. 102

[217] Zhai, Y., Tehrani, P., Li, L., Zhao, J., and Zhao, Q. (2011). Dynamic pricing under binary demand uncertainty: A multi-armed bandit with correlated arms, *Proc. of the 45th IEEE Asilomar Conference on Signals, Systems, and Computers*. DOI: 10.1109/acssc.2011.6190289. 124, 125

[218] Zhao, Q., Krishnamachari, B., and Liu, K. (2008). On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance, *IEEE Transactions on Wireless Communications*, 7(12):5431–5440. DOI: 10.1109/t-wc.2008.071349. 119

[219] Zhao, Q. and Sadler, B. (2007). A survey of dynamic spectrum access, *IEEE Signal Processing Magazine*, 24(3):79–89. DOI: 10.1109/msp.2007.361604. 117

[220] Zoghi, M., Whiteson, S., Munos, R., and de Rijke, M. (2014). Relative upper confidence bound for the k-armed dueling bandit problem, *Proc. of International Conference on Machine Learning (ICML)*, pages 10–18. 103

[221] Zoghi, M., Karnin, Z. S., Whiteson, S., and de Rijke, M. (2015). Copeland dueling bandits, *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pages 307–315. 103

Author's Biography

QING ZHAO

Qing Zhao is a Joseph C. Ford Professor of Engineering at Cornell University. Prior to that, she was a Professor in the ECE Department at University of California, Davis. She received a Ph.D. in Electrical Engineering from Cornell in 2001. Her research interests include sequential decision theory, stochastic optimization, machine learning, and algorithmic theory with applications in infrastructure, communications, and social-economic networks. She is a Fellow of IEEE, a Distinguished Lecturer of the IEEE Signal Processing Society, a Marie Skłodowska-Curie Fellow of the European Union Research and Innovation program, and a Jubilee Chair Professor of Chalmers University during her 2018–2019 sabbatical leave. She received the 2010 IEEE Signal Processing Magazine Best Paper Award and the 2000 Young Author Best Paper Award from IEEE Signal Processing Society.