

On Thompson Sampling with Noninformative Priors  
in Stochastic Multi-Armed Bandits  
(確率的多腕バンディットにおける無情報事前分布を用いた  
トンプソン抽出について)

by

Jongyeong Lee  
李 鍾瑛

A Doctor Thesis  
博士論文

Submitted to  
the Graduate School of the University of Tokyo  
on June 7, 2023  
in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Information Science and Technology  
in Computer Science

Thesis Supervisor: Masashi Sugiyama 杉山 将  
Professor of Computer Science

## ABSTRACT

We human beings always make a decision on a daily basis. Such a decision would be a trivial one, such as what to eat for lunch, or an important business one that determines the future of the company. This is why decision-making modeling has been proposed and analyzed in several research contexts, such as business, political psychology, statistics, and computer science. In particular decision-making scenarios, one can observe all possible outcomes regardless of the decisions they made. However, making decisions is not always straightforward, as we often have to take action based on imperfect observations. This is because we often observe only the consequences of our actions, which is known as *partial feedback*.

In the sequential scenario, an agent is faced with the challenge of gathering information about different actions while minimizing the negative impact of making poor decisions. They can achieve this by balancing the exploration of unknown actions to reduce uncertainty and the exploitation of well-known actions that are expected to be effective. The *Multi-Armed Bandit (MAB) problem* is a classical model that exemplifies such a trade-off between exploration and exploitation in sequential decision-making. In the stochastic setting studied in this thesis, the agent observes an outcome (or reward) every time they select an action (or arm), which is generated from the underlying arm distribution specific to that arm. To solve such problems efficiently, the agent would pose an assumption on the reward distribution with their prior knowledge. For example, one could assume that rewards are generated from a Bernoulli distribution if rewards are binary.

In stochastic MAB problems, two objectives have been mainly considered: (1) *regret minimization* where the agent aims to maximize their rewards during the time horizon, which is a problem of balancing the trade-off between exploration and exploitation. (2) *best-arm identification* where the agent aims to identify the best (optimal) arm with limited resources, which is a problem of pure exploration.

*Thompson sampling* is a randomized probability matching policy that solves the stochastic MAB problem by playing arms according to the posterior probability of being optimal. From its randomization nature, Thompson sampling naturally makes the balance between exploration and exploitation and has shown excellent performance in practice. However, despite its effectiveness, Thompson sampling was not considered an attractive choice because of its lack of theoretical understanding for a long time. In the last decade, the theoretical analyses of Thompson sampling have been conducted, and now it is considered one of the main approaches to the stochastic MAB problem.

In much literature on Thompson sampling in the regret minimization problem, the underlying distribution was assumed to belong to the single parameter exponential family, such as the Bernoulli distribution, or to be a light-tailed distribution, where its tail probability decreases exponentially. This is because the sub-Gaussian noise is widely observable in many problems, and it is easy to control the probability of extreme events. In practice, however, multi-parameter or heavy-tailed distributions have also been widely adopted to analyze stochastic systems in several research fields, such as economics and signal processing. Nevertheless, Thompson sampling has rarely been investigated under such distributions, even in the parametric stochastic MAB problem, which is a classical problem.

The first part of this thesis aims to deepen the theoretical understanding of Thompson sampling in such distributions from a problem-dependent view, where we show that the problem-dependent optimality of Thompson sampling depends on the choice of (non-informative) priors. More precisely, we first provide a theoretical analysis of Thompson sampling in the uniform distribution with unknown support, which is a two-dimensional parametric distribution and does not belong to the exponential family. We further propose a variant of Thompson sampling that could maintain optimality under one-to-one reparameterization, as our analysis shows the optimality of Thompson sampling further depends on the parameterization of distributions when one employs the uniform priors. We then extend our analysis to the Pareto distribution, which is a heavy-tailed distribution parameterized by two unknown parameters. It is worth noting that online ad allocation is one main application of bandit algorithms, and the Pareto distribution is commonly observed in the analysis of the internet and web.

The second part of this thesis focuses on the best-arm identification problem with Thompson sampling. Despite its efficient exploration in the regret minimization problem, direct use of Thompson sampling in pure exploration cannot make the optimal algorithm. This is because

Thompson sampling plays a suboptimal arm a log-order times, which induces a bias in the number of playing arms. To address this challenge, we concentrate our efforts on leveraging the inherent randomization of Thompson Sampling. By integrating Thompson sampling and deterministic algorithms, we could construct a more computationally efficient exploration strategy.

In summary, this thesis aims to extend the theoretical understanding of Thompson sampling in stochastic MAB problems, placing a specific emphasis on guiding the selection of priors in general models. Such extensions are not only for mathematical interests, but they will also be helpful to practitioners who want to solve sequential decision-making problems through the application of easy-to-implement algorithms that guarantee excellent performance in both theory and practice. The insights and findings presented in this thesis significantly contribute to our understanding of Thompson Sampling and provide enhanced and efficient solutions to MAB problems.

## Acknowledgements

This thesis has been accomplished under the supervision of Professor Masashi Sugiyama, who has been my mentor throughout my graduate journey. I would like to express my deepest gratitude to him for his persistent guidance, invaluable research opportunities, and the great academic environment he provided. These factors have significantly influenced my exploration and understanding of this expansive field of study. His unwavering support has been crucial in shaping my academic journey and establishing my research interests. Additionally, I would like to extend my deepest gratitude to Professor Junya Honda, whose guidance has been indispensable in developing my understanding of bandit problems and improving my technical writing skills throughout my Ph.D. journey. Without his invaluable assistance, the completion of this thesis would not have been possible.

I would also like to express my sincere appreciation to my thesis committee: Professor Ken-ichi Kawarabayashi, Professor Yusuke Miyao, Professor Akiko Aizawa, Professor Tetsuo Shibuya, and Professor Akiko Takeda. Their willingness to devote their time and expertise to the assessment of my work has provided me with an invaluable opportunity to enhance the quality of my thesis. I am deeply grateful for their insightful feedback and constructive criticisms, which have undoubtedly enriched my research and deepened my understanding of the subject matter.

Moreover, I am extremely thankful to Professor Chao-Kai Chiang, whose insightful advice and lively discussions have been a source of continual inspiration. His guidance has not only enriched my research work but also instilled confidence in my abilities. Furthermore, I am deeply thankful to Nontawat Charoenphakdee, who made my transition into this new environment smoother and showed me great care and support.

I also want to extend my gratitude to all the current and past members of our laboratory for creating an enjoyable and productive atmosphere. I am particularly thankful to my lab mates Masahiro Fujisawa (who kindly answered my questions on Bayesian stuff), Tianyi Zhang, Yivan Zhang (we had an amazing time in Honolulu), and Tsuchiya Taira (who studied the bandit problems together). Also, I would like to thank Johannes Ackermann and Riou Charles, with whom I spent almost the entire period in the lab during the COVID era. Since I have always been interested in other cultures, especially food, talking with them was a valuable experience to learn about cultural differences and similarities. I would like to thank Zhenghang Cui, Cemal Erat, Anan Methasate, and Meike Tütken for playing board games and sharing their cultures with me.

Next, I would like to thank my previous lab mates Takuya Shimada (who invited me to his weddings!), Takuo Kaneko, Hideaki Imamura, and Yutaka Kitamura. I enjoyed talking with them, which also made my transition into this new environment smoother. I would like to thank my senior colleagues Yuko Kuroki, Seiichi Kuroki, Han Bao, Takeshi Teshima, Kento Nozawa, Masahiro Kato and other lab mates Shintaro Nakamura, Zhenguo Wu, Xinqiang Cai, Valliappa Chockalingam, Thanawat Lodkaew, Yuting Tang, Xiaoyu Dong, Xiaomou Hou, Iu Yahiro, Kazuki Ota, Yu Yao, Tiankui Xian, Masahiro Negishi, and Yuto Nozaki.

Finally, I want to express my gratitude to my parents and sisters. Their unconditional



support has allowed me to focus on my studies and explore my interests without distractions. Their encouragement and faith in me have been the backbone of my academic journey.

I was supported by Support for Pioneering Research Initiated by the Next Generation (SPRING) Program of the Japan Science and Technology Agency (JST), JST SPRING, Grant Number JPMJSP2108, October 2021 to March 2023.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Multi-Armed Bandit Problem . . . . .	1
1.1.1	Variety in reward generation . . . . .	2
1.1.2	Variety in learning objective . . . . .	3
1.1.3	Variety in strategies . . . . .	4
1.1.4	Applications . . . . .	7
1.2	Thesis Overview . . . . .	8
1.2.1	Challenges and motivations . . . . .	8
1.2.2	Thesis objective . . . . .	9
1.3	Thesis Organization . . . . .	10
1.4	Thesis Contribution . . . . .	12
1.4.1	Analysis of TS in the model of non-regular multiparameter distributions . . . . .	12
1.4.2	Development of a variant of TS with a simple design principle . . . . .	12
1.4.3	Application of TS as an exploration tool in the BAI problems . . . . .	13
<b>2</b>	<b>Preliminaries</b>	<b>15</b>
2.1	Formulation of Multi-Armed Bandit . . . . .	15
2.1.1	Regret . . . . .	16
2.1.2	Best-arm identification . . . . .	18
2.2	Policies for Stochastic MAB . . . . .	20
2.2.1	Upper confidence bound approach . . . . .	20
2.2.2	Thompson sampling . . . . .	21
2.3	Thompson Sampling and Priors . . . . .	21
2.3.1	Conjugate prior . . . . .	22
2.3.2	Noninformative prior . . . . .	22
2.3.3	Performance of Thompson sampling . . . . .	29
<b>3</b>	<b>Uniform Bandits</b>	<b>30</b>
3.1	Introduction . . . . .	30
3.1.1	Chapter background . . . . .	30
3.1.2	Chapter contribution . . . . .	31
3.1.3	Chapter organization . . . . .	32
3.2	Problem Formulation . . . . .	32
3.3	Thompson Sampling and the Choice of Priors . . . . .	33
3.3.1	Thompson sampling for the uniform bandits . . . . .	34
3.3.2	Thompson sampling with truncation for the LS family . . . . .	34
3.4	Main Theoretical Results . . . . .	36
3.5	Gaussian Bandits . . . . .	38
3.5.1	Thompson sampling for the Gaussian bandits . . . . .	39
3.5.2	Thompson sampling with truncation for the Gaussian bandits . . . . .	39
3.6	Simulation Results . . . . .	40

3.7	Proofs of Theoretical Results . . . . .	42
3.7.1	Derivation of the posteriors for the uniform model . . . . .	42
3.7.2	Proof of main results . . . . .	42
3.7.3	Proof of Lemma 3.8 . . . . .	44
3.7.4	Proof of Lemma 3.9 . . . . .	45
3.7.5	Proofs of technical lemmas for Lemma 3.8 . . . . .	52
3.7.6	Proof of Lemma 3.10 . . . . .	55
3.7.7	Proof of Lemma 3.11 . . . . .	57
3.7.8	Proofs of technical lemmas for Lemma 3.11 . . . . .	63
3.7.9	Proof of the suboptimality of TS . . . . .	70
3.8	Conclusion . . . . .	73
<b>4</b>	<b>Pareto Bandits</b>	<b>74</b>
4.1	Introduction . . . . .	74
4.1.1	Chapter background . . . . .	74
4.1.2	Chapter contribution . . . . .	75
4.1.3	Chapter organization . . . . .	76
4.2	Problem Formulation . . . . .	76
4.2.1	Pareto distribution . . . . .	76
4.2.2	Asymptotic regret lower bound . . . . .	77
4.2.3	Relation with bounded moment models . . . . .	78
4.3	Thompson Sampling and the Choice of Priors . . . . .	78
4.3.1	TS and TS-T for the Pareto bandits . . . . .	79
4.3.2	Interpretation of the prior parameter $k$ . . . . .	80
4.4	Main Theoretical Results . . . . .	81
4.5	Simulation Results . . . . .	83
4.6	Proofs of Theoretical Results . . . . .	85
4.6.1	Closed form of the problem-dependent constant . . . . .	85
4.6.2	Derivation of the posteriors . . . . .	89
4.6.3	Proof of the optimality of TS and TS-T . . . . .	90
4.6.4	Proof of Lemma 4.5 . . . . .	90
4.6.5	Proof of Lemma 4.6 . . . . .	91
4.6.6	Proofs of technical lemmas for Lemma 4.5 . . . . .	92
4.6.7	Proofs of technical lemmas for Lemma 4.6 . . . . .	100
4.6.8	Proofs of technical lemmas on fundamental inequalities . . . . .	107
4.6.9	Proof of suboptimality of TS . . . . .	110
4.6.10	Concentration inequalities . . . . .	114
4.7	Conclusion . . . . .	114
<b>5</b>	<b>Thompson Exploration</b>	<b>116</b>
5.1	Introduction . . . . .	116
5.1.1	Chapter background . . . . .	116
5.1.2	Chapter contribution . . . . .	117
5.1.3	Chapter organization . . . . .	118
5.2	Problem Formulation . . . . .	118
5.2.1	Notation and SPEF bandits . . . . .	118
5.2.2	Stopping rule . . . . .	120
5.3	Best Challenger with Thompson Exploration . . . . .	120
5.4	Main Theoretical Results . . . . .	122
5.4.1	Main theorems . . . . .	122
5.4.2	Comparison with $\beta$ -optimality . . . . .	123
5.4.3	Comparison with asymptotic optimality . . . . .	124

5.5	Simulation Results . . . . .	127
5.6	Proofs of Theoretical Results . . . . .	128
5.6.1	Proof of Lemma 5.2 . . . . .	130
5.6.2	Proof of Theorem 5.3: Convergence of estimates . . . . .	131
5.6.3	Proofs of technical lemmas for Theorem 5.3: Sufficient conditions for the convergence of estimates . . . . .	136
5.6.4	Proofs of technical lemmas for Theorem 5.3: Boundedness of the number of rounds where estimates do not converge . . . . .	137
5.6.5	Proof of technical lemma for Theorem 5.3: An upper bound on the number of rounds where TE occurs . . . . .	140
5.6.6	Proof of technical lemma for Theorem 5.3: Analysis with TS . . . . .	142
5.6.7	Proof of technical lemma for Lemma 5.8 . . . . .	143
5.6.8	Proof of technical lemma for Lemma 5.12 . . . . .	143
5.6.9	Proof of Theorem 5.4: Sample complexity . . . . .	149
5.6.10	Proof of Lemma 5.19 . . . . .	155
5.7	Conclusion . . . . .	156
<b>6</b>	<b>Conclusions and Future Work . . . . .</b>	<b>158</b>
6.1	Summary of the Thesis . . . . .	158
6.2	Future Directions . . . . .	159
6.2.1	Minimax optimality and asymptotic optimality . . . . .	160
6.2.2	Model misidentification . . . . .	161
6.2.3	Theoretical derivation of the relationship between TS and priors . . . . .	162
	<b>References . . . . .</b>	<b>163</b>

## List of Figures

1.1	Structure of this thesis. Regret minimization and best-arm identification indicate the learning objective considered in the following chapters . . .	10
1.2	An example where the posterior distribution of each arm belongs to the Gaussian distribution. The solid lines represent the posterior probability of sampling mean values, while the blue and red dashed lines indicate the <i>true</i> expected value of each arm, respectively. . . . .	13
3.1	A two-armed example where the posterior of $\mu$ is given as $\text{Uniform}_{\mu\sigma}(\hat{\mu}, \hat{\sigma})$ and TS-T replaces $\hat{\sigma}$ with $\bar{\sigma} = \max(\hat{\sigma}, 0.5)$ . Suppose that $\hat{\mu}_B = 1 < \hat{\mu}_A = 1.2$ holds in some rounds. The shaded regions denote the probability of the currently suboptimal arm B being played when (a) $\tilde{\mu}_A = 1.05$ and (b) $\tilde{\mu}_A = 0.9$ are denoted by a star mark. Different line widths are used to distinguish two lines when they are overlapped. . . . .	35
3.2	Cumulative regret for the 6-armed uniform bandit instance $\nu_6^U$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively. . . . .	40
3.3	Cumulative regret for the 6-armed Gaussian bandit instance $\nu_6^G$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively. . . . .	41
4.1	Cumulative regret for the 4-armed Pareto bandit instance $\nu_4^{(1)}$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively. The blacked dotted line denotes the problem-dependent lower bound based on Lemma 4.1. . . . .	83
4.2	The solid lines denote an averaged regret over independent 10,000 runs for the 4-armed Pareto bandit instance $\nu_4^{(1)}$ . The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of the regret, respectively. . . . .	84
4.3	Cumulative regret for the 4-armed Pareto bandit instance $\nu_4^{(2)}$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively. The green dotted line denotes the problem-dependent lower bound based on Lemma 4.1. . . . .	85
4.4	The solid lines denote an averaged regret over independent 10,000 runs for the 4-armed Pareto bandit instance $\nu_4^{(2)}$ . The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of the regret, respectively. . . . .	85

5.1	Stopping times of various policies for 5-armed Bernoulli bandit instance with means $\mu_5^B = (0.3, 0.21, 0.2, 0.19, 0.18)$ and different maximal risk over 3,000 independent runs. The black star denotes the mean of stopping times. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a]. . . . .	129
5.2	Stopping times of various policies for 4-armed Gaussian bandit instance with means $\mu_4^G = (1.0, 0.85, 0.8, 0.7)$ and unit scale and different maximal risk over 3,000 independent runs. The black star denotes the mean of stopping times. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a].	129
6.1	The relationship between the difficulty of an instance and the regret. Recall the definition of the uniformly fast convergent (UFC) policy in Definition 2.1, where Lai and Robbins [1985] and Burnetas and Katehakis [1996] showed that any UFC policies could not be below the “asymptotic optimal” line at any point. Note that it is not possible for any policy to be completely below the “minimax optimal” line. This illustration is inspired by Figure 16.1 in Lattimore and Szepesvári [2020]. . . . .	161

## List of Tables

2.1	Asymptotic optimality of Thompson sampling with different noninformative priors for various models. ✓ and ✗ denote whether TS can achieve the asymptotic lower regret bound in (2.1) for the corresponding rewards model or not. . . . .	29
3.1	Parameters of the 6-armed bandit instances. . . . .	41
4.1	Parameters of the 4-armed bandit instances. . . . .	83
5.1	Sample complexity for 5-armed Bernoulli bandit instance with means $\mu^B = (0.3, 0.21, 0.2, 0.19, 0.18)$ over 3,000 independent runs. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a]. . . . .	128
5.2	Sample complexity for 4-armed Gaussian bandit instance with means $\mu_4^G = (1.0, 0.85, 0.8, 0.7)$ with unit scale over 3,000 independent runs. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a]. . . . .	128
5.3	Average time of one step of various policies. . . . .	128
6.1	Asymptotic optimality of TS with different noninformative priors for multiparameter models. R, C, and T denote whether the model satisfies the Fisher information regularity (✓) or not (✗), whether it is compact (✓) or non-compact (✗), and whether its function is light-tailed (L) or heavy-tailed (H). O-T and O-TT denote whether TS and TS-T can achieve the asymptotic lower regret bound in (2.1) for the corresponding model (✓) or not (✗), respectively. Notice that $_H$ in O-T indicates that the results are derived by Honda and Takemura [2014]. $\pi_u$ , $\pi_j$ , and $\pi_r$ denote the uniform prior for the specific parameterization in superscript, the Jeffreys prior, and the reference priors, respectively. . . . .	160

# Chapter 1

## Introduction

In today’s fast-paced world, we are constantly faced with choices at every turn. Whether we are choosing which restaurants to eat for lunch, buying which board games to play on Saturday night, or deciding on which platforms to place an advertisement for new products, each of these situations requires us to make a choice in order to progress. Some choices are particularly challenging because we must navigate an uncertain environment where the outcomes of our decisions are not entirely clear. In other words, life is a complex web of decisions, ranging from mundane daily decisions to complex, high-stakes dilemmas, and our ability to navigate this web is a testament to our adaptability and resourcefulness.

*Sequential decision-making under uncertainty* is a field of study that aims to understand and optimize the process of making decisions in uncertain environments. The roots of sequential decision-making can be traced back to the 19th century when occasional studies were conducted on hypothesis testing [Kotz and Johnson, 2012]. Nonetheless, sequential decision-making was more thoroughly developed in the mid-20th century, especially in the fields of statistics [Neyman and Pearson, 1933, Wald, 1945], economics [Simon, 1959], and psychology [Becker, 1958, Edwards, 1954].

With this long history, the field has evolved to encompass a wide range of topics, such as Markov decision processes [Howard, 1960, White, 1993], reinforcement learning [Sugiyama, 2015, Sutton and Barto, 1998], active learning [Bryson and Mobolurin, 1997, Rubens et al., 2015], to name a few. Although various methods and techniques have been developed, a fundamental aspect of sequential decision-making under uncertainty is the trade-off between exploration and exploitation. This trade-off encapsulates the dilemma of deciding when to gather more information (exploration) and when to use the information already acquired to make the best possible decision (exploitation).

### 1.1 Multi-Armed Bandit Problem

This thesis focuses on the *multi-armed bandit (MAB) problem*, a fundamental sequential decision-making model that exemplifies the exploration-exploitation dilemma. The problem is named after the term for slot machines, “one-armed bandit,” and involves a gambler or agent who must choose among multiple options, or “arms,” each characterized by an *unknown* probability of providing a reward. With limited knowledge of these probabilities, the agent has to choose an arm carefully to maximize rewards based on the history of their choices and corresponding rewards. The difficulty of this problem originated from “partial feedback,” where the agent is only able to observe a reward from the played arm. Therefore, balancing between playing the seemingly best arm and obtaining information about other arms becomes crucial in order to maximize cumulative rewards.

The idea of the bandit problem was initially introduced by Thompson [1933] in the context of clinical treatment as a two-armed bandit problem. Despite the significance of



this early contribution, it remained largely unrecognized for a considerable period. Two decades later, Robbins [1952] independently formalized the MAB problem. Since then, it has become increasingly popular, particularly following the seminal works by Lai and Robbins [1985] and Burnetas and Katehakis [1996], where they provided critical insights into the structure of the problem, performance bounds, and optimal strategies. These theoretical developments offered a deeper understanding of the exploration-exploitation trade-off, which in turn motivated several research and applications [Auer et al., 2002, Bouneffouf et al., 2020, Preil and Krapp, 2022]. Another key advantage of the MAB problem is its simplicity in comparison to reinforcement learning or Markov decision processes, as it does not involve state transitions or actions that impact the environment [Lattimore and Szepesvári, 2020, Sutton and Barto, 1998]. This simplicity allows researchers and practitioners to delve into the fundamental understanding inherent in exploration and exploitation, which are at the heart of all sequential decision-making models.

### 1.1.1 Variety in reward generation

In the MAB problem, a finite set of arms is available to the agent, and each arm is associated with a specific reward distribution. Various adaptations of the MAB problem have been introduced in the literature to reflect different assumptions on reward distributions [Auer et al., 2002, Bubeck and Cesa-Bianchi, 2012, Komiyama et al., 2015, Zimmert and Seldin, 2021].

In the foundational *stochastic bandit* problem, the reward distribution of each arm is assumed to be fixed, but its parameter remains unknown to the agent. Whenever the agent plays an arm, they observe feedback generated from the corresponding reward distribution, which is assumed to be independent and identically distributed. Therefore, the exploration-exploitation dilemma can be represented by the balancing between playing the arm that, based on the accumulated rewards, currently seems to be the best (exploitation) and probing other arms to gain more information on their unknown parameter (exploration). In this classical stochastic setting, the reward distributions are assumed to belong to a certain statistical model. For example, Lai and Robbins [1985] considered the specific families indexed by a single parameter.

On the other hand, in a non-stationary bandit scenario, the reward distributions are no longer fixed [Gittins, 1979, Whittle, 1988]. As noted in Lattimore and Szepesvári [2020, section 31.3], such time-varying environments can also be modeled by the contextual bandit problem, where the agent receives additional contextual information before each action [Langford and Zhang, 2007, Luo et al., 2018]. This contextual information is typically referred to as the state of the environment, which can influence the reward distribution of each arm. It is worth noting that the action of the agent does not change the state in the non-stationary bandit problem, which distinguishes it from reinforcement learning [Sugiyama, 2015, Sutton and Barto, 1998].

Such assumptions on the reward distribution are eliminated in the adversarial bandit [Auer et al., 1995, Uchiya et al., 2010, Wei and Luo, 2018]. In this generalization of the stochastic bandit problem, rewards can even be non-stochastic. Instead, an adversary determines the rewards for each arm at every step, possibly in a way that is adversarial to the agent’s policy. Therefore, the adversarial bandit problem can be seen as a two-player zero-sum game between the agent and the adversary, where a randomized policy is necessary for the agent to avoid being exploited by the adversary since the agent is always the first player.

Other variants of the bandit problem include the dueling bandit [Jamieson and Nowak, 2011, Yue et al., 2012], the combinatorial bandit [Chen et al., 2013, Gai et al., 2010], and bandit with graph feedback [Chen et al., 2021, Liu et al., 2018, Mannor and Shamir,

2011], to name a few. Despite these variants, this thesis focuses on the stochastic bandit problem, which allows us to concentrate on the *fundamental* issues inherent in the exploration and exploitation dilemma without being distracted by additional complexities.

### 1.1.2 Variety in learning objective

Not only the diversity in the reward generation models but several learning objectives also have been considered to handle different learning situations. In this thesis, we roughly divided it into two categories: regret minimization and pure exploration.

The exploration-exploitation dilemma is rooted in the desire of an agent to maximize rewards in an uncertain environment. This leads to the concept of regret, defined as the opportunity cost or the loss induced by playing a suboptimal arm, which can be formalized as the difference between the rewards of the optimal arm and the played arm. Therefore, *minimizing regret* is essentially equivalent to maximizing reward, which is the first objective of the bandit problem [Auer et al., 1995, Burnetas and Katehakis, 1996, Lai and Robbins, 1985, Robbins, 1952]. To be more precise, the objective of the regret minimization problem is to find a policy that makes an exquisite balance between exploration and exploitation, thereby minimizing the cumulative regret (total loss) incurred by their choices. Since regret is the core notion in the bandit problems, several variants have been introduced, such as Bayesian regret [Hong et al., 2022, Kveton et al., 2021, Russo and Van Roy, 2014], minimax regret [Audibert and Bubeck, 2009, Cai and Pu, 2022, Li et al., 2019], and approximation regret [Chen et al., 2013, Kakade et al., 2007, Streeter and Golovin, 2008].

Although various types of policies have been proposed to address the problem of regret minimization in stochastic settings, two approaches have become prevalent in the literature. The first approach is the upper confidence bound (UCB) policy, which computes a high-probability confidence bound on the reward of each arm [Auer et al., 2002, Lai and Robbins, 1985]. The UCB policy follows the principle of optimism in the face of uncertainty, where the arm with the largest upper confidence bound is selected [Abbasi-Yadkori et al., 2011, Kaufmann et al., 2012a]. It is worth noting that the UCB policies are deterministic with respect to the accumulated observations, as the confidence bound of each arm is determined solely by the observed rewards. The second approach is Thompson sampling (TS), which is known as the first policy in the bandit problem [Thompson, 1933]. Unlike the UCB policy, TS is a randomized Bayesian policy that plays an arm according to the posterior probability of being optimal [Chapelle and Li, 2011, Russo et al., 2018]. Both approaches have gained popularity for their excellent empirical performance supported by strong theoretical guarantees [Bubeck and Cesa-Bianchi, 2012].

However, in certain scenarios where the exploration and exploitation phases are separated, the focus may be solely on exploration without worrying about any induced regret in the exploration phase [Bubeck et al., 2011, Kalyanakrishnan et al., 2012, Mannor and Tsitsiklis, 2004]. For example, one can consider the development of a new drug, where the researchers would aim to identify the most effective treatment from a set of alternatives before being tested on a large group of patients. In such cases, Bubeck et al. [2009] showed that the policies designed to achieve logarithmic regret in regret minimization could not perform satisfactorily in the pure exploration problem, which stimulates researchers to take a different approach.

A common challenge in such settings is the task of identifying the optimal arm, known as the *best arm identification* (BAI) problem, where an agent aims to find the arm with the largest reward within a limited budget of resources [Even-Dar et al., 2002, 2006, Maron and Moore, 1997]. Several policies have been proposed to address the BAI problem under various assumptions about reward generation, e.g., the canonical

exponential family [Garivier and Kaufmann, 2016], the linear bandit [Soare et al., 2014, Xu et al., 2018], and the combinatorial bandit [Chen et al., 2014, Du et al., 2021]. Note that the two-armed BAI problem can be seen as A/B testing [Dimakopoulou et al., 2021, Kaufmann et al., 2014].

In addition to the BAI problem, the bandit literature has introduced other variants of pure exploration. One such variant is the threshold bandits, which involve identifying all arms whose rewards exceed a certain threshold [Cheshire et al., 2021, Locatelli et al., 2016]. Another variant is stochastic optimization, which can be seen as a continuous-armed bandit problem [Dani et al., 2008, Lazaric and Brunskill, 2014]. More details and other variants can be found in the bibliographical remarks in Lattimore and Szepesvári [see 2020, Section 33.5].

### 1.1.3 Variety in strategies

To address the diversity in bandit problems, various policies have been proposed. While the details of each policy may differ, some of them share common design principles, allowing them to be categorized into a comprehensive framework.

#### Explore-then-commit

The explore-then-commit (ETC) strategy is one of the earliest policies proposed in the context of the MAB problems [Anscombe, 1963, Maurice, 1957, Robbins, 1952, Somerville, 1954], which consists of two distinct phases: an exploration phase and an exploitation phase. During the exploration phase, the ETC strategy plays every arm alternatively a certain number of times to gather information. Then, it simply plays the arm with the largest observed expected reward in the exploration phase for the remaining rounds.

One of the key advantages of the ETC strategy is its simplicity and ease of implementation, as it does not require complex computations or assumptions about the underlying reward distribution when the exploration sample size is determined. The ETC strategy has not only been utilized in the past, but has also been used recently to solve context bandit [Langford and Zhang, 2007], linear bandit [Rusmevichientong and Tsitsiklis, 2010], and batch bandit [Perchet et al., 2016]. In the context of two-armed bandit problems, Garivier et al. [2016] theoretically showed that the ETC strategy can outperform a UCB-based policy when the gap between the expected rewards of the two arms is known. However, when the gap is unknown, which would be the usual case in practice, they proved that the ETC strategy is necessarily suboptimal when the objective is minimizing regret.

Additionally, there are also relatives of the ETC strategy. The elimination strategy maintains a set of active arms and sequentially eliminates arms based on observations [Even-Dar et al., 2006, Shahrampour et al., 2017], which can be seen as a generalization of the ETC strategy to the general MAB problems. However, similar to the ETC strategy, the elimination-based policy is also shown to be suboptimal [see Lattimore and Szepesvári, 2020, Exercise 6.8].

Another well-known variant would be the  $\epsilon$ -greedy policy that plays an arm uniformly random with probability  $\epsilon$  and plays the currently best arm with probability  $1 - \epsilon$ , which is a randomized variant of the ETC strategy [Auer et al., 2002, Sutton and Barto, 1998]. For the bandit problems, Auer et al. [2002] showed that  $\epsilon$ -greedy policy is suboptimal although it performs well in practice if one chooses  $\epsilon$  appropriately. Apart from the bandit problems, this randomized strategy is popular and widely used in reinforcement learning due to its simplicity and generality [Dabney et al., 2021, Kalashnikov et al., 2018, Mnih et al., 2015] even though it requires an exponential sample complexity

to learn value functions in reinforcement learning, which is suboptimal [Osband et al., 2019].

### **Upper confidence bound**

The UCB strategy is widely recognized as one of the most mainstream approaches to solving stochastic MAB problems due to its adaptability to various scenarios [Agrawal, 1995, Auer et al., 2002, Garivier and Cappé, 2011, Jamieson et al., 2014, Katehakis and Robbins, 1995, Lai, 1987, Srinivas et al., 2010, Takemura et al., 2021]. In the UCB strategy, each arm is evaluated based on an index that is defined as the sum of the empirical mean estimates and the confidence width which is an exploration bonus. Therefore, this index represents the upper limit of the expected reward with a certain level of confidence.

By constructing confidence intervals with consistent estimators of the mean of reward distributions, the confidence interval narrows as more observations are gathered, leading to a reduction in the exploration bonus. This dynamic adjustment of the exploration bonus allows the UCB strategy to effectively allocate exploration efforts in the early stages and gradually shift towards exploiting the currently best arms as more observations become available. As a result, the UCB strategy makes a balance between exploration and exploitation through its adaptive exploration bonus, contributing to its theoretical optimality in various scenarios [Garivier and Cappé, 2011, Garivier et al., 2022, Jamieson et al., 2014].

The adaptability of the UCB strategy is not limited to the settings where the interests lie in the mean rewards. For instance, Sani et al. [2012] addressed risk-aversion in bandit problems by considering both the mean and variance of reward distributions, and Galichet et al. [2013] tackled conditional-value-at-risk maximization problems by considering the lower limit of the expected rewards. A more systematic approach to handling different risk-related criteria with UCB-based policies in stochastic MAB problems was proposed by Cassel et al. [2018], further highlighting the flexibility and wide applicability of the UCB strategy in diverse MAB settings.

While the maximum likelihood estimators can serve as a consistent estimator of mean rewards in many cases, the choice of the confidence level is a delicate and challenging task, as it directly impacts the exploration bonus. A too-small confidence level may fail to motivate exploration of currently suboptimal arms, while a too-large confidence level may cause an excessive play of suboptimal arms, resulting in large regret. Nevertheless, various choices of confidence levels have been considered to address specific problems optimally, demonstrating its adaptability to various settings [Kaufmann et al., 2012a, Ramamohan et al., 2016].

### **Thompson sampling**

As briefly introduced in Section 1.1.2, TS is a Bayesian policy that maintains the posterior distribution of the rewards instead of constructing a confidence interval that is used in frequentist statistics [Thompson, 1933]. In TS, the agent observes samples generated from the current posterior distribution at every round and simply plays the arm with the largest expected rewards computed based on these posterior samples. This is contrast to the UCB strategy, where the index is deterministic to the accumulated observations. The inherent randomness in TS through posterior distributions enables it to automatically make a balance between exploration and exploitation, leading to excellent performance in numerous applications [Chapelle and Li, 2011, May et al., 2012, Russo et al., 2018, Scott, 2010]. This outstanding empirical performance is attributed to a remarkable resurgence of interest in the past decade although TS, which is the oldest heuristic bandit algorithm, remained largely unnoticed for a significant period. As a result, TS

has gained significant attention from researchers, leading to extensive theoretical analyses that have deepened our understanding of its properties and performance [Agrawal and Goyal, 2012, 2013, Honda and Takemura, 2014, Kaufmann et al., 2012b, Riou and Honda, 2020]. This combination of empirical success and rigorous theoretical analysis has made TS a prominent and valuable tool for solving various MAB problems.

### **Main interest in this thesis**

In addition to the above strategies, there are other strategies such as the information directed sampling [Hao et al., 2022, Kirschner et al., 2020, 2021, Russo and Van Roy, 2014] and the minimum empirical divergence strategy [Honda and Takemura, 2010, 2011, Komiyama et al., 2015, Saber et al., 2021] in regret minimization, and the racing rule [Garivier and Kaufmann, 2016, Grover et al., 2018, Kaufmann and Kalyan Krishnan, 2013] and the top-two sampling [Jourdan et al., 2022, Qin et al., 2017, Russo, 2016, Shang et al., 2020] in best arm identification. Nevertheless, this thesis focuses on the analysis of TS due to its generality, simplicity, adaptability, and outstanding empirical performance. The challenges of TS and the details on TS are given in Sections 1.2 and 2.2, respectively.

**Generality** As long as an agent has access to a posterior distribution, regardless of how it was derived or the complexity of the underlying model, TS can be implemented and utilized effectively. The generality of TS, coupled with its adaptability elucidated below, has made it a popular and widely used strategy in practice, even in scenarios where strong theoretical guarantees are neither always available nor necessary [Chapelle and Li, 2011, Scott, 2010].

**Simplicity** The simplicity of TS is one of its key advantages, contributing to its popularity and widespread use. As explained above, TS follows the simple two-step process: it observes samples from the posterior distributions of each arm and selects the arm with the largest sampled expected reward. This sample-and-select approach is straightforward to implement and computationally efficient, making it attractive to both researchers and practitioners [see Russo et al., 2018, and references therein]. Although the ETC strategy is also very simple and efficient, its suboptimality is widely known in general [Auer et al., 2002]. On the other hand, the simplicity of TS, combined with its excellent empirical performance and theoretical guarantees, has made it a popular choice in many decision-making scenarios [Baudry et al., 2021, Ferreira et al., 2018].

**Adaptability** Similarly to the UCB strategy, TS can also handle various settings such as conditional-value-at-risk maximization [Baudry et al., 2021] and mean-variance minimization [Zhu and Tan, 2020]. A unified analysis of TS for continuous risk-averse bandit problems was provided by Chang and Tan [2022], demonstrating its applicability in addressing risk-related criteria. Furthermore, its flexibility enables us to incorporate prior knowledge on the arms through prior distributions [Bubeck and Cesa-Bianchi, 2012]. Its inherent randomness plays a key role in handling uncertainty in MAB problems, contributing to its outstanding empirical performance, even though it makes the analysis complicated. This randomness also makes TS robust to delayed feedback [Chapelle and Li, 2011], providing distinct advantages over deterministic strategies such as the UCB strategy and the ETC strategy. In Chapter 5, we further employ the inherent randomness of TS to tackle the BAI problem, further highlighting its adaptability and effectiveness in handling diverse settings.

### 1.1.4 Applications

Here, we briefly introduce a few applications of the bandit algorithms.

#### Online advertising

When advertising companies place an advertisement on the web, they aim to maximize revenue by carefully selecting the most suitable advertisement for each user. In such situations, the reward can be defined by the likelihood of clicking on the advertisement, which is commonly referred to as the click-through rate. However, due to budget constraints, the agent must balance between allocating new advertisements for exploration and exploiting the currently most effective ones. Therefore, the MAB framework has primarily been applied in online advertising [Pandey and Olston, 2006, Schwartz et al., 2017, Xu et al., 2013], with applications extending to recommendation systems [Brodén et al., 2017, Zhou et al., 2017] and A/B testing [Degenne et al., 2019b, Kaufmann et al., 2014].

Various extensions to the traditional MAB problem have been proposed to capture additional complexities in real-world scenarios. For example, the common challenges in online advertising such as delayed feedback [Chapelle, 2014, Gael et al., 2020, Pike-Burke et al., 2018] and budget constraints [Avadhanula et al., 2021, Badanidiyuru et al., 2018] have been considered. Furthermore, MAB algorithms can be used for personalized advertising by leveraging contextual information to display different advertisements to users based on their individual preferences and behaviors [Han and Gabor, 2020, Nuara et al., 2018, Tang et al., 2015]. These developments have led to significant improvements in the efficacy and efficiency of online advertising, highlighting the value of the MAB framework as a tool in digital marketing [Aramayo et al., 2023, Cao and Sun, 2019].

#### Sequential experimental design

Similarly to the bandit problems, experimental design has been considered in the area of medical trials [Anscombe, 1963, Armitage, 1960], where each trial is often expensive and can be potentially harmful to the subject. Since the objective is efficiently identifying the optimal treatment while minimizing the number of trials and associated risks, the MAB framework is well-suited for modeling such sequential decision-making problems in experimental design [Hardwick and Stout, 1991, Vogel, 1960, Zhang and Lee, 2010]. For a more comprehensive survey on the use of the bandit framework in experimental design, refer to Burtini et al. [2015].

#### Dynamic pricing

In dynamic pricing, an agent continuously adjusts the price of their product in response to varying market conditions, such as changes in supply and demand [Den Boer, 2015, Elmaghraby and Keskinocak, 2003]. Since the agent only observes whether a customer bought the product without knowing the maximum price each customer is willing to pay, the MAB framework becomes useful for maximizing revenue through real-time price adjustments based on customer responses [Babaioff et al., 2012, Wang, 2007]. A significant advantage of employing MAB algorithms in this context is their ability to handle intricate pricing scenarios, such as high-dimensionality [Mueller et al., 2019, Roth et al., 2020] and multimodality [Wang et al., 2021b]. As a result, MAB-based dynamic pricing strategies have been shown to outperform conventional policies [Misra et al., 2019].

## Other applications

The MAB framework has also been applied to various problems where one needs to make a sequential decision in the presence of uncertainty regarding the outcome. For instance, Baudry et al. [2021] applied TS to decide the planting date in agriculture, and Preil and Krapp [2022] recently introduced the MAB frameworks in the field of inventory optimization, to name a few. For a more detailed overview of how the bandit framework is being applied across various fields, readers may refer to Bouneffouf et al. [2012].

## 1.2 Thesis Overview

This section aims to provide a concise introduction to the background and motivation of this thesis, outlining the key research questions and challenges. By doing so, we clarify the context of this thesis, which helps the readers understand the significance and relevance of the study. In addition, this outline offers a glimpse into the following chapters, giving readers a preview of the topics that will be discussed in detail.

### 1.2.1 Challenges and motivations

The simplicity in the formulation of the MAB problem has led to a strong focus on the theoretical understanding of sequential decision-making models and the analysis of policies in the bandit literature [Burnetas and Katehakis, 1996, Lai and Robbins, 1985]. As illustrated in Section 1.1.4, this emphasis on theoretical foundations has motivated researchers and practitioners to apply these mathematically well-established bandit policies to their own problem domains.

However, it is important to note that some bandit policies are specifically designed to achieve optimal theoretical performance for a specific problem, which can come at the cost of high computational complexity, e.g., solving optimization problems at every round [Agrawal et al., 2021b, Garivier and Kaufmann, 2016]. As a result, researchers have also proposed various heuristics that are computationally efficient or show excellent performance in practice, even if they lack theoretical guarantees [Henderson et al., 2018, Xia and Yap, 2018]. It is worth noting that these heuristics may not be applicable or suitable for other problem domains, as they are designed for specific problems and lack theoretical guarantees.

In other words, to overcome the challenges of bridging the gap between theory and practice in the sequential decision-making model, it is necessary to propose a comprehensive framework that combines theoretical guarantees with practical efficiency. The main challenges that guide this thesis are summarized as follows.

Developing a general framework that not only provides theoretical guarantees but is also computationally efficient can greatly benefit the field of sequential decision-making. Furthermore, this framework should be designed to be easily implementable and have the potential to serve as a valuable guideline for solving various real-world problems.

As briefly introduced in Section 1.1.2, two types of bandit policies, namely the UCB policy and TS, stand out in the various problem domains [Gopalan and Mannor, 2015, Talebi et al., 2017, Yoon and Chow, 2020, Zenati et al., 2022]. It is worth noting that there has been a remarkable resurgence of interest in TS in the past decade, even though

it is the oldest heuristic bandit algorithm that remained largely unnoticed for a significant period. This renewed attention can be attributed to its outstanding empirical performance, which has been demonstrated in numerous successful applications [Chapelle and Li, 2011, May et al., 2012, Russo et al., 2018, Scott, 2010]. Following its empirical success, TS has gained significant attention from researchers, leading to extensive theoretical analyses that have deepened our understanding of its properties and performance [Agrawal and Goyal, 2012, 2013, Kaufmann et al., 2012b].

Nevertheless, the effectiveness of the UCB and TS approaches heavily relies on the design of the confidence bounds and the choice of prior distributions, respectively. Note that these tasks are not straightforward and may depend on the problem at hand. For example, determining the significance level for constructing confidence bounds in UCB is challenging, as it directly impacts the exploration-exploitation trade-off of the policy [Lattimore and Szepesvári, 2020]. Similarly, in TS, the choice of priors significantly influences its performance since the sampling is based on the posterior distribution [Russo et al., 2018]. However, selecting appropriate priors is a complex task, even in the context of inference problems in Bayesian analysis [Robert, 2007].

In this thesis, we evaluate the performance of TS in terms of the expected regret, although the evaluation of TS often relies on the Bayesian regret, which serves as a Bayesian counterpart to the expected regret [Bubeck and Sellke, 2020, Dubey and Pentland, 2019, Russo and Van Roy, 2016]. However, Bayesian regret is not always directly interpretable in terms of the expected regret. On the other hand, the upper bound of the expected regret of TS can provide an upper bound on the Bayesian regret, although bounding the expected regret of TS is usually technically complicated [Lattimore and Szepesvári, 2020].

Currently, most theoretical analyses of the expected regret of TS focus on specific priors that yield optimal performance in particular problem settings [Kaufmann et al., 2012b, Korda et al., 2013, Riou and Honda, 2020]. Although these optimal results may be sufficient for those specific problems, their generalizability to other problems remains uncertain. Furthermore, the impact of prior selection on the performance of TS in bandit problems remains unclear, limiting our ability to guess the performance of TS when applied to new problems.

### 1.2.2 Thesis objective

Discussion in Sections 1.1.3 and 1.2.1 highlighted several key points regarding TS:

- TS is a general framework that can be applied to several sequential decision-making models under uncertainty.
- Implementation of TS is straightforward and computationally efficient, particularly when closed-form posterior distributions can be obtained.
- Both theoretical analyses and empirical evaluations have demonstrated the outstanding performance of TS.

However, there is a missing piece in TS to address the main challenge highlighted in Section 1.2.1, particularly when dealing with new models that lack theoretical understanding. This leads to the central question that has guided this thesis:

Is there a universally applicable prior in general decision-making models that consistently leads to high-performance outcomes when employed in TS?



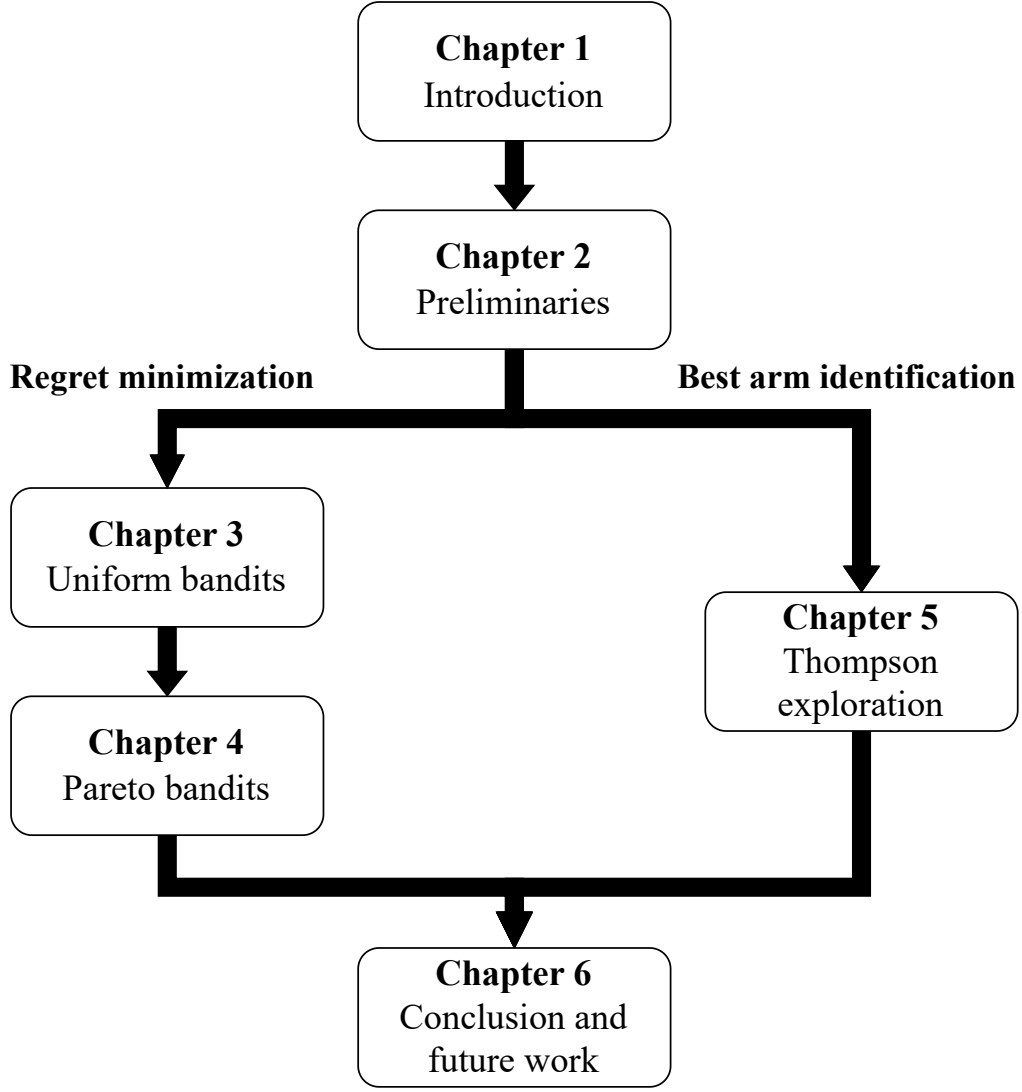


Figure 1.1: Structure of this thesis. Regret minimization and best-arm identification indicate the learning objective considered in the following chapters

In general, there are no silver bullets that can solve all problems optimally. However, it might be able to discover a “bronze bullet”, a solution that can solve some problems optimally and also performs reasonably well in others. With this in mind, the author believes that this thesis serves as a stepping stone for future applications and analysis of other variants. By deepening our understanding of the role of priors in stochastic bandit problems, this study enhances our ability to tackle future research and applications in this field. It provides a solid foundation upon which further investigations can be built, paving the way for new insights and developments in the field of sequential decision-making.

### 1.3 Thesis Organization

In this section, we provide an outline of the thesis and a brief summary of each chapter. The entire structure of this thesis is illustrated in Figure 1.1.

**Chapter 2: Preliminaries** This chapter starts by introducing the notation that will be used throughout the thesis. This notation allows us to precisely define the key concepts

in the bandit problems, which were previously described verbally in Chapter 1. Then, we discuss relevant notions of optimality and formulate two major approaches: the UCB policy and TS. Given the main interest of the thesis on TS, we introduce several well-known priors that have been extensively studied in Bayesian analysis. Finally, we end this chapter by presenting the result of the existing literature to examine the impact of priors on the performance of TS in bandit problems.

**Chapter 3: Uniform bandits** The existing analysis demonstrated that TS could achieve optimal performance for the model of the single parameter exponential family (SPEF) by selecting natural priors such as the uniform prior and the Jeffreys prior [Kaufmann et al., 2012b, Korda et al., 2013]. As a result, it is widely recognized that the performance of TS is not significantly affected by the choice of priors unless a widely adopted prior is selected. This is because, in the bandit problem, only the expectation of the reward distribution is of interest, whereas in certain inference problems, various parameterizations are of interest, making the choice of priors more crucial. However, Honda and Takemura [2014] showed that TS with the reference prior and the Jeffreys prior are suboptimal when considering the Gaussian model with unknown mean and variance, while TS with the uniform prior achieves optimal performance under location-scale parameterization.

In this chapter, we first show that the prior sensitivity of TS occurs not only in the noncompact multiparameter models but also in the uniform model with unknown supports, which is a compact non-regular multiparameter model. Furthermore, we explicitly formulate the suboptimality of the uniform prior under different parameterizations, including the location-rate parameterization, not only for uniform models but also for Gaussian models. These results undermine the trustworthiness of the uniform prior in different models, particularly when reparameterization is considered for efficient computations. To address these issues, we propose a variant of TS with a general design idea that can achieve optimal performance with the reference prior and the Jeffreys prior for both the uniform models and the Gaussian models. Additionally, we provide numerical results that support our theoretical findings.

**Chapter 4: Pareto bandits** In this chapter, the main focus is on the non-regular heavy-tailed distribution, the Pareto distribution, which is commonly used in the analysis of social science [Mahanti et al., 2013], such as the number of visitors to a Web site [Clauset et al., 2007]. By modeling the number of website visitors using a Pareto distribution, we can explore the applicability of the Pareto bandit model in online advertising, which is a main application of bandit policies. The Pareto distribution differs notably from the models discussed in the previous chapter in two aspects: (i) it exhibits a heavy tail and may have infinite variance, and (ii) its expected value cannot be defined for certain parameters.

In this context, we first prove the suboptimality of TS with both the reference prior and the Jeffreys prior. Interestingly, we also found that using a uniform prior not only makes the implementation of TS complicated but also leads to suboptimal performance under the scale-shape parameterization, which is a natural parameterization in Pareto models, similar to location-scale in Gaussian models. To be precise, this uniform prior performs better than the Jeffreys prior but worse than the reference prior. These results further undermine the reliability of the uniform prior as a “default” prior for general models. Moreover, we extend our investigation to solve the suboptimality of the reference priors and the Jeffreys prior using the same technique employed in the previous chapter. Based on these results, we conjecture that the reference prior, along with a pre-processed version of TS, can serve as a reference choice when one considers a new model since the reference posterior can be defined universally and is invariant if the

reparameterization does not change the group order of parameters.

**Chapter 5: Thompson exploration** In this chapter, we consider the BAI problem, where we are only interested in the quality of a final decision rather than the performance of the test phase. Since a direct application of the policy for regret minimization performs suboptimally in this problem [Bubeck et al., 2009], one needs to design a specific policy for the BAI problems. However, some of them are computationally heavy as they solved an optimization problem at every round [Korda et al., 2013], or are forced to play an arm at least a certain number of times [Ménard, 2019, Wang et al., 2021a]. To address these limitations, we propose a novel policy in this chapter that combines TS with a heuristic approach, resulting in a computationally efficient method that naturally explores without the need for forced exploration steps. We show that our proposed method is asymptotically optimal for any two-armed bandit problems and achieves near optimality for general MAB problems. We highlight the advantages of our near optimality by comparing it with the concept of  $\beta$ -optimality, which is commonly used in the analysis of Bayesian policies [Qin et al., 2017, Russo, 2016, Shang et al., 2020]. It is worth noting that the theoretical analysis of the simple heuristic algorithm was previously unknown, making our analysis presented in this chapter particularly interesting and valuable.

**Chapter 6: Conclusion and future work** This chapter concludes this thesis by summarizing the findings we discovered. Then, we discuss potential future directions in investigating the reliability and generalizability of our “bronze bullet”, the pre-processed reference posterior matching policy, in more practical situations and different evaluation metrics.

## 1.4 Thesis Contribution

In this section, we summarize the key contributions made in this thesis, which can be categorized into three main topics.

### 1.4.1 Analysis of TS in the model of non-regular multiparameter distributions

This thesis contributes to the theoretical understanding of TS in the context of multiparameter distributions, which has received relatively less attention despite its practical usefulness. We first show that the prior sensitivity of TS occurs not only in the non-compact multiparameter models but also in the uniform model with unknown supports, which is a compact non-regular multiparameter model. Furthermore, we observe consistent results in the Pareto model, a noncompact non-regular heavy-tailed multiparameter model. Specifically, the Jeffreys priors suffer from a polynomial regret for certain bandit instances. While TS with the uniform prior demonstrates optimal performance in Gaussian and uniform models on the location-scale parameterization, we prove its suboptimality in different parameterizations. These findings emphasize the significance of selecting appropriate priors for regret minimization in multiparameter models, which poses a nontrivial challenge due to the absence of a universal solution.

### 1.4.2 Development of a variant of TS with a simple design principle

We propose a variant of TS, TS with truncation (TS-T), where we devise an adaptive truncation procedure on the parameter space of the posterior distribution to control the problems in the early stage of learning. The underlying design principle of TS-T can be described in one sentence as follows:

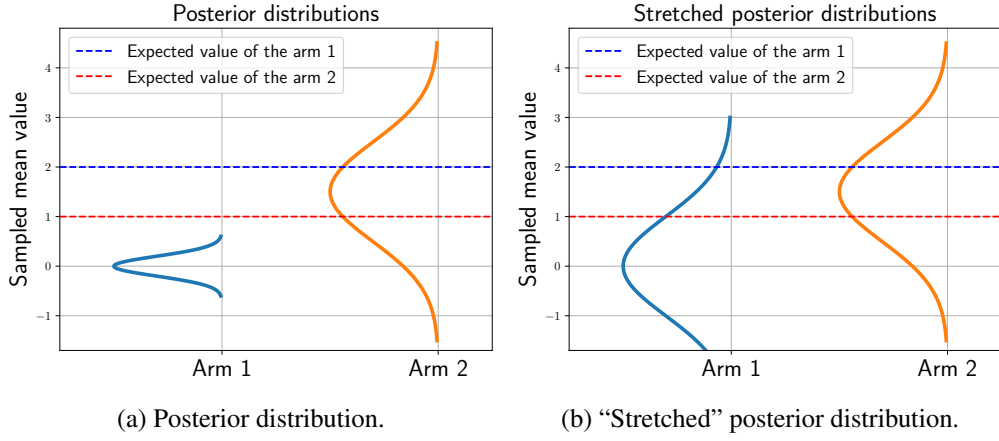


Figure 1.2: An example where the posterior distribution of each arm belongs to the Gaussian distribution. The solid lines represent the posterior probability of sampling mean values, while the blue and red dashed lines indicate the *true* expected value of each arm, respectively.

Truncate the parameter space of the posterior distribution to stretch the distribution, which encourages a policy to explore in the early stage of learning.

With this simple idea, we prove the optimality of TS with the reference prior and the Jeffreys prior for all models considered in this thesis. This offers an alternative approach to achieving optimality without the need to search for an optimal or appropriate prior for each specific problem, a process we expect will be significantly more convenient in practical applications.

**An illustrative example: stretching the posterior distribution** Here, we present a simple example to illustrate the design principle proposed in this thesis. We consider the case where the posterior distribution is given as a Gaussian distribution, as shown in Figure 1.2.

Figure 1.2a shows the posterior distribution of each arm based on the observed rewards. In the early stage of learning, due to its randomness, the posterior distribution of the optimal arm (arm 1) is concentrated on the small value, 0, in this example. This can lead to suboptimal behavior, where TS is more likely to play arm 2, which is expected to generate a higher reward according to the posterior probability.

To address this issue, we introduce truncation to the parameter space of the posterior distribution, where we lift the scale parameter of the Gaussian distribution in Figure 1.2b, which helps prevent extreme cases from occurring in the early stage of learning. Here, stretching the posterior distribution can be seen as flattening the posterior distributions, which prevents them from overly concentrating on the specific value in the early stage of learning. By flattening the distributions, we encourage exploration and avoid prematurely favoring a specific arm based on limited observations. It is important to design the truncation carefully to cover the entire parameter space of the posterior distribution as the number of samples increases. Note that this example highlights the effectiveness of truncation in mitigating suboptimal behavior during the learning process. Further details and analysis can be found in the subsequent chapters of this thesis.

### 1.4.3 Application of TS as an exploration tool in the BAI problems

We propose a computationally efficient policy for the BAI problems, which avoids solving optimization problems at each round by computing a subgradient instead. While

the use of subgradients has been explored in previous research [Ménard, 2019, Wang et al., 2021a], our contribution lies in adapting TS with the Jeffreys prior<sup>1</sup> as an exploration tool that eliminates the need for an artificial forced exploration step. Our proposed policy can be applied to the single-parameter exponential family, while the analysis of some existing works is limited to the Gaussian models or the Bernoulli models. Additionally, our analysis has its own interest since we analyze the heuristic algorithm whose optimality was previously unknown and demonstrate that our proposed policy achieves optimality up to a constant factor. Notably, our policy naturally incorporates exploration, distinguishing it from the other Bayesian policies that play the currently best arm with predefined probability.

---

<sup>1</sup>Note that it is equivalent to the reference prior for the single parameter exponential family [Ghosh, 2011], which is considered in Chapter 5.

## Chapter 2

### Preliminaries

In the previous chapter, we introduced the stochastic multi-armed bandit (MAB) problem and its variety in two major criteria. In this chapter, we provide a comprehensive review of related topics. We begin by presenting a formal definition of the stochastic MAB problem and formulating the key concepts and terminology that are essential to understand this study. Building upon these foundations, we review two outstanding approaches, the upper confidence bound policy and Thompson sampling.

**Chapter Organization** The organization of this chapter is as follows. In Section 2.1, we introduce the notations used throughout this thesis and use them to formalize the stochastic MAB problem along with its objectives and related optimality notions. Section 2.2 provides a brief overview of two well-known policies that have been proposed to solve the stochastic MAB problem. We provide the details of Thompson sampling, which is the main topic of this thesis in Section 2.3.

#### 2.1 Formulation of Multi-Armed Bandit

In this section, we provide general notations that are used in the following chapters and formulate the MAB problem and related key concepts.

**Set** We denote the set of real, integral, and natural numbers by  $\mathbb{R}$ ,  $\mathbb{Z}$ , and  $\mathbb{N}$ , respectively. We use the subscript with inequality to denote the subset of numbers that satisfies the condition, for example,  $\mathbb{Z}_{\geq 2} := \{z \in \mathbb{Z} : z \geq 2\}$ . For simplicity, we sometimes use  $\mathbb{R}_+$  and  $\mathbb{R}_-$  instead of  $\mathbb{R}_{>0}$  and  $\mathbb{R}_{<0}$ . The probability simplex is denoted by  $\Sigma_K = \{\mathbf{w} \in [0, 1]^K : \sum_{i=1}^K w_i = 1\}$ . For an integer  $n$ , we denote  $[n] := \{1, 2, \dots, n\}$  instead of  $\mathbb{N}_{\leq n}$  for simplicity.

**Stochastic MAB** For  $K \in \mathbb{N}_{\geq 2}$ , we consider  $K$ -armed bandit problems, where an agent plays an arm  $i \in [K]$  at each round  $t \in \mathbb{N}$ . The generated reward of each arm  $i$  is independent and identically distributed (i.i.d.) from its corresponding probability distribution  $\nu_i = \nu_{\theta_i}$ , parameterized by  $\theta_i \in \mathbb{R}^d$  for  $d \in \mathbb{N}$ . We denote the  $n$ -th observation from the arm  $i$  by  $X_{i,n} \stackrel{\text{i.i.d.}}{\sim} \nu_i$ . Let  $N_i(t) = \sum_{s=1}^{t-1} \mathbb{1}[i(s) = i]$  be the number of rounds that the arm  $i$  is played until round  $t$ , where  $\mathbb{1}[\cdot]$  denotes the indicator function. In this setting, the agent observes the reward only from the played arm  $i(t)$  at time  $t$ .

In the stochastic MAB problem, it is typically assumed that the distribution  $\nu$  is known, but its associated parameter  $\theta$  is unknown to the agent. For instance, in the Gaussian bandits, the reward distributions are Gaussian distributions with unknown location and scale parameters. Let  $\mu(\theta) = \mathbb{E}_{\nu_\theta}[X] = \int_{-\infty}^{\infty} x d\nu_\theta(x)$  denote the expected value of  $\nu_\theta$ , which we call the expected reward or the mean reward interchangeably. Let

$\mu_i = \mu(\theta_i)$  denote the expected reward of the arm  $i \in [K]$ , and we denote the maximum reward by  $\mu^* = \max_{i \in [K]} \mu_i$ .

### 2.1.1 Regret

As we introduced in Section 1.1.2, one of the main objectives in the MAB problem is to maximize the total reward obtained over a certain time horizon  $T \in \mathbb{N}$ . However, due to the inherent uncertainty caused by the unknown parameters of the reward distribution, it is often difficult to determine which arm yields the largest expected reward. Therefore, it is necessary to introduce a quantitative measure that captures the difference between the agent's policy and the optimal policy that plays the optimal arm at all rounds. This measure is commonly known as *regret*, which reflects the cumulative cost (loss) of playing the arm  $i(t)$  at each round  $t$ . Then, pseudo-regret, or simply, regret in this thesis, at round  $T$  is defined with the sub-optimality gap  $\Delta_i := \mu^* - \mu_i$  as

$$\text{Reg}(T) = \text{Reg}(T; \Pi, \nu) := T\mu^* - \sum_{t=1}^T \mu_{i(t)} = \sum_{t=1}^T \Delta_{i(t)},$$

where  $\Pi$  denotes the policy and  $\nu = (\nu_{\theta_i})_{i=1}^K$  denotes bandit instance. Although regret depends on both the policy and the bandit instance, we drop its dependence throughout this thesis unless it is not obvious. Since regret is a fundamental concept in bandit problems, we will introduce several variants of regret as follows.

#### Expected regret

By taking expectations with respect to the measure on outcome induced by the policy and bandit instances, we define the expected regret as

$$\mathbb{E}[\text{Reg}(T; \Pi, \nu)] = T\mu^* - \mathbb{E} \left[ \sum_{t=1}^T \mu_{i(t)} \right].$$

Therefore, the expected regret measures the regret of a particular policy on a particular bandit instance, where the sub-optimality gap  $\Delta$  is regarded as a fixed quantity.

This problem-dependent regret takes into account the inherent difficulty of the specific bandit instance, which can vary depending on the number of arms, their distributions, and the suboptimality gap. While designing a policy for a specific bandit instance can be successful, it may not generalize well to other instances. Therefore, one needs to design a policy that can perform well for any instance in a *reasonable* manner, which requires the policy not to be overly specialized for certain instances. Such a concept was formalized by Lai and Robbins [1985] and Burnetas and Katehakis [1996] in the following definition.

**Definition 2.1.** For a class of bandit instances  $\mathcal{P} := \{(\nu_{\theta_i})_{i=1}^K : \forall i \in [K], \theta_i \in \Theta_i\}$  where  $\Theta_i \in \mathbb{R}^{d_i}$  is a known parameter set for a dimension  $d_i \in \mathbb{N}$  of  $\theta_i$ , a policy  $\Pi$  is called *uniformly fast convergent* over  $\mathcal{P}$ , if for all  $\nu \in \mathcal{P}$  and  $\alpha > 0$  it holds that

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T; \Pi, \nu)]}{n^\alpha} = 0.$$

This means that any uniformly fast convergent policy does not incur a polynomial expected regret for all instances of interest. Furthermore, they provided the asymptotic regret lower bound that any uniformly fast convergent policy must satisfy

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \geq \sum_{i: \mu_i < \mu^*} \frac{\Delta_i}{\inf_{\theta \in \Theta_i: \mu(\theta) > \mu^*} \text{KL}(\nu_{\theta_i}; \nu_\theta)}, \quad (2.1)$$

where  $\text{KL}(\cdot; \cdot)$  denotes the Kullback-Leibler (KL) divergence between probability distributions. A policy  $\Pi$  is called *asymptotically optimal* policy on class  $\mathcal{P}$  when it satisfies for any  $\nu \in \mathcal{P}$  that

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T; \Pi, \nu)]}{\log T} \leq \sum_{i: \mu_i < \mu^*} \frac{\Delta_i}{\inf_{\theta \in \Theta_i: \mu(\theta) > \mu^*} \text{KL}(\nu_{\theta_i}; \nu_{\theta})}.$$

### Worst-case regret

The expected regret is used to evaluate the performance of a policy given a specific instance  $\nu$  in the class of bandit instances  $\mathcal{P}$ . In contrast, the worst-case regret provides an upper bound on the expected regret in the worst-case scenario, which is defined by

$$\text{WorstReg}(T; \Pi, \mathcal{P}) = \sup_{\nu \in \mathcal{P}} \mathbb{E}[\text{Reg}(T; \pi, \nu)].$$

Here, one says that a policy  $\Pi$  has no regret on the class  $\mathcal{P}$  if the policy eventually plays the optimal arm most of the time, which is formulated as

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \text{WorstReg}(T; \Pi, \mathcal{P}) \rightarrow 0.$$

The worst-case regret shows the performance of a policy under the most challenging problem and thus is useful to evaluate the robustness of a policy. Therefore, it has been widely used to analyze the high-probability performance of policies for the Markov decision process [Agrawal et al., 2021a, Agrawal and Jia, 2017, Jaksch et al., 2010].

### Minimax regret

While the regrets introduced above evaluate the performance of a specific policy, the minimax regret evaluates the inherent complexity of bandit problems, focusing on the problem itself rather than individual policies. It is defined by the lowest worst-case regret across the set of policies  $\Pi$ , which is written as

$$\text{MinimaxReg}(T; \Pi, \mathcal{P}) = \inf_{\Pi \in \Pi} \text{WorstReg}(T; \Pi, \mathcal{P}).$$

Here, a policy that achieves the minimax regret is said to be minimax optimal [Audibert and Bubeck, 2009, Jin et al., 2021].

### Bayesian regret

Bayesian regret is another variant of regret that has been used to evaluate the performance of Bayesian policies [Bubeck and Sellke, 2020, Russo and Van Roy, 2016], which is defined by the expected regret over all problem instances with respect to the prior  $\pi$ ,

$$\text{BayesReg}(T; \Pi, \pi) = \mathbb{E}_{\theta \sim \pi} [\mathbb{E}[\text{Reg}(T; \Pi, \nu_{\theta})]].$$

However, the Bayesian regret of a policy can be less informative in the sense that a bound on the Bayesian regret could provide a loose bound on the problem-dependent regret in general [Lattimore and Szepesvári, 2020]. Notice that an upper bound on the expected regret can provide information to the Bayesian regret from its definition.



## Main interest in this thesis

So far, we have introduced several regrets considered in bandit literature, which are developed from different perspectives. The interest of this thesis lies in the expected regret of a policy, and we aim to build an asymptotically optimal policy in Chapters 3 and 4. This is because the worst-case regret is often too conservative to provide a meaningful bound in practice since it sometimes becomes excessively large due to an extremely difficult instance. Although the minimax regret can provide a deeper understanding of the complexity of the problem, finding a minimax optimal policy is computationally expensive in practice, which makes researchers focus on developing nearly minimax optimal policies recently [Lee et al., 2021, Li et al., 2019, Zhou et al., 2021].

### 2.1.2 Best-arm identification

Apart from the regret minimization problem, the objective in the best-arm identification (BAI) problems is to identify the optimal arm that has the largest expected regret  $\mu^*$  given limited resources [Bubeck et al., 2011, Gabillon et al., 2012]. In general, any policies to solve the BAI problem have the following structure [Garivier and Kaufmann, 2016].

- A *sampling rule*: decides which arm  $i(t)$  to be played at every round based on past observations.
- A *stopping rule*: determines when the policy stops the sampling procedure.
- A *decision rule*: outputs which arm is possibly the optimal arm when the policy stops.

Based on the types of limited resources, two variants of the BAI problem have been mainly studied so far.

#### Fixed budget

When the number of possible trials  $\tau$ , which we call the budget, is predetermined, the objective of the agent becomes identifying the optimal arm as *accurately* as possible for the given budget. Therefore, the stopping rule in this setting can be described as “stop after  $\tau$  rounds,” and the objective can be formulated by

$$\text{for given } \tau \in \mathbb{N} : \quad \text{minimize } \mathbb{P}_\nu [\mu_{\text{output}} \neq \mu^*],$$

where output denotes the final decision made by the policy and  $\nu$  denotes the bandit instance of interest. In this setting, a policy is said to be *consistent* if it satisfies for any bandit instances  $\nu$  belonging to a class  $\mathcal{P}$

$$\limsup_{T \rightarrow \infty} \mathbb{P}_\nu [\mu_{\text{output}} \neq \mu^*] = 0.$$

For a two-armed bandit model with  $\mu_1 > \mu_2$ , Kaufmann et al. [2016] provided a lower bound on the probability of misidentification,

$$\begin{aligned} \limsup_{\tau \rightarrow \infty} -\frac{1}{t} \log \mathbb{P}_\nu [\mu_{\text{output}} \neq \mu^*] \\ \leq \inf_{\nu_{\theta'_1}, \nu_{\theta'_2} \in \mathcal{P} : \mu(\theta'_1) < \mu(\theta'_2)} \max \left( \text{KL} \left( \nu_{\theta'_1}; \nu_{\theta_1} \right), \text{KL} \left( \nu_{\theta'_2}; \nu_{\theta_2} \right) \right), \end{aligned}$$

which is later improved by Carpentier and Locatelli [2016]. In this setting, the probability of misidentification decays exponentially for some rates, and hence one aims to build an algorithm that decays at the optimal rate [Komiyama et al., 2022].

## Fixed confidence

In the fixed confidence setting, the agent aims to build a policy whose probability of misidentification is less than the predetermined risk parameter  $\delta \in (0, 1)$ . In other words, they aim to design a  $\delta$  probably approximately correct (PAC) policy, which satisfies for  $\nu \in \mathcal{P}$

$$\mathbb{P}_\nu [\mu_{\text{output}} \neq \mu^*] \leq \delta.$$

Let  $\tau_\delta$  be the stopping time of a policy with risk  $\delta$ , which denotes the number of rounds until the policy stops. Then, their objective is to identify the optimal arm with accuracy at least  $1 - \delta$  as *soon* as possible, i.e., as small  $\tau_\delta$  as possible.

When  $\mathcal{P}_e$  denotes a set of exponential bandit models where every instance  $\nu \in \mathcal{P}_e$  has a unique optimal arm, there exists an arm  $i^*(\nu)$  such that  $\mu(\theta_{i^*(\nu)}) > \max_{i \neq i^*(\nu)} \mu_i$ . Here, Garivier and Kaufmann [2016] showed that any  $\delta$ -PAC policy satisfies for any  $\nu \in \mathcal{P}_e$  and  $\delta \in (0, 1)$ ,

$$\mathbb{E}_\nu[\tau_\delta] \geq T^*(\nu) \log \left( \frac{1}{2.4\delta} \right), \quad (2.2)$$

where

$$T^*(\nu) := \left( \sup_{\mathbf{w} \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) \right)^{-1} \quad (2.3)$$

and  $\text{Alt}(\nu) := \{\lambda \in \mathcal{P}_e : i^*(\lambda) \neq i^*(\nu)\}$ . Here, the solution of (2.3) is denoted by  $\mathbf{w}^*(\nu)$ , which indicates the optimal proportion of arm plays to achieve the lower bound in (2.2). Similarly to the regret lower bound in (2.1), Garivier and Kaufmann [2016] provided the problem-dependent asymptotic lower bound on the sample complexity  $[E][\tau_\delta]$ , which is

$$\liminf_{\delta \rightarrow 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \geq T^*(\nu).$$

A policy satisfying

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T^*(\nu)$$

is said to be *asymptotically optimal* in the fixed confidence setting [Barrier et al., 2022, Degenne et al., 2019a, 2020].

As a relaxed notion of the asymptotic optimality, the  $\beta$ -optimality has been considered in Bayesian sampling rules [Jourdan et al., 2022, Qin et al., 2017, Russo, 2016, Shang et al., 2020]. Specifically, a policy is said to be asymptotically  $\beta$ -optimal if it satisfies

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log(1/\delta)} \leq T^\beta(\nu),$$

where

$$T^\beta(\nu) := \left( \sup_{\mathbf{w} \in \Sigma_K, w_{i^*(\nu)} = \beta} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) \right)^{-1}.$$

From its definition,  $T^*(\nu) = \min_{\beta \in [0, 1]} T^\beta(\nu)$  holds. Thus, the  $\beta$ -optimality does not necessarily imply the optimality in the sense of the tight lower bound in (2.2) unless optimal  $\beta$  is equal to  $w_{i^*(\nu)}^*(\nu)$ .

## Main interest in this thesis

A fixed budget setting would be useful when the agent has a very strict resource limitation to pay, while a fixed confidence setting is useful to obtain a reliable decision with affordable accuracy. As noted in Lattimore and Szepesvári [2020], the constraint

on the horizon makes the problem complex and the results are not as straightforward compared to the fixed confidence. Therefore, several asymptotically optimal algorithms have been proposed in the fixed confidence setting [Garivier and Kaufmann, 2016, Ménard, 2019, Wang et al., 2021a]. However, most of them are equipped with the *forced exploration* steps, which are designed to explore suboptimal arms sufficiently to achieve asymptotic optimality. Since the forcing exploration would induce a problem especially when the number of arms is large, we aim to avoid it in Chapter 5 by using a randomized approach.

## 2.2 Policies for Stochastic MAB

With the various perspectives and objectives introduced so far, a large number of policies have been proposed to address the diversity in the field. Although the details of a policy can be different, some of them can be gathered through the high-level strategies they employ to solve problems. Here, we present some major approaches that have been developed to solve the problems of interest in this thesis.

### 2.2.1 Upper confidence bound approach

The seminal work by Lai and Robbins [1985] provided not only the regret lower bound but also the idea of constructing an asymptotically optimal policy, based on the sample-from-the-leader principle. They defined the leader at round  $t$  for some predetermined  $\delta \in (0, 1)$  as

$$\text{leader} = \arg \max_{i \in [K]: N_i(t) \geq \delta t} \hat{\mu}_{i, N_i(t)},$$

where  $\hat{\mu}_{i, n}$  denotes the empirical mean of the arm  $i$  after observing  $n$  generated rewards. The threshold  $\delta t$  is used to ensure that the empirical mean accurately estimates the true mean reward. Therefore, leader is the current best arm among the arms with a well-estimated empirical mean. The balance between exploration and exploitation is controlled by comparing the upper confidence bound (UCB) of each arm, which is the upper limit of the confidence interval. By designing a significance level, one can make the confidence width decrease, which encourages exploring an arm sufficiently to reduce the uncertainty. In short, the UCB-based policy is a kind of index policy that constructs a UCB index  $\text{UCB}_i(\delta)$  for each arm  $i$  with significance level  $\delta \in (0, 1)$ , which satisfies

$$\mathbb{P}[\mu_i > \text{UCB}_i(\delta)] \leq \delta.$$

Although one can choose any  $\delta \in (0, 1)$ , it is a very delicate problem since it controls the balance between exploration and exploitation, which determines the performance of a policy.

The design principle in the UCB approach has been known as *optimism in the face of uncertainty*, which suggests that one should act as if the best outcome will occur. Based on this optimistic principle, many versions of UCB policies have been proposed to solve different bandit instances [Abbasi-Yadkori et al., 2011, Garivier and Cappé, 2011, Kaufmann et al., 2012a]. For instance, Auer et al. [2002] proposed the UCB1 policy for the bounded support bandits  $\mathcal{P} = \{\nu : \text{supp}(\nu_i) \in [0, 1], \forall i \in [K]\}$ . At each round  $t \in \mathbb{N}$ , they set the significance level  $\delta = \frac{1}{t}$  and UCB index as

$$\text{UCB}_i(\delta) = \hat{\mu}_{i, N_i(t)} + \sqrt{\frac{\log t}{2N_i(t)}}.$$

From Hoeffding's inequality, one can check that

$$\mathbb{P}[\mu_i \geq \text{UCB}_i(\delta)] \leq e^{-2N_i(t)(\text{UCB}_i(\delta) - \hat{\mu}_{i, N_i(t)})^2} = \frac{1}{N_i(t)}.$$

---

**Algorithm 1** Thompson Sampling

---

- 1: **Parameters:** Set prior distribution  $\pi_0$ .
  - 2: **for**  $t = 1, \dots, T$  **do**
  - 3: Sample  $\tilde{\theta}_i(t) \sim \pi(\theta_i \mid X_{i,N_i(t)})$  for each  $i \in [K]$ .
  - 4: Play  $i(t) = \arg \max_{i \in [K]} \mu(\tilde{\theta}_i)$ .
  - 5: Observe a reward  $x_{i(t), N_{i(t)}(t)+1} \sim \nu_{\theta_{i(t)}}$ .
- 

However, the choice of the significance level is not a trivial task, and the design of the UCB index for each problem typically requires theoretical analysis to ensure its performance. Furthermore, since the index is deterministic to past observations, it might be overly exploratory or not exploratory enough due to the randomness inherent in the stochastic bandit problems.

### 2.2.2 Thompson sampling

Thompson sampling (TS) is a randomized Bayesian policy based on the probability matching principle, which selects an arm according to the *posterior* probability of the arm being optimal at each round. As the number of samples increases, the posterior distributions of each arm become more concentrated around certain values, resulting in a decrease in the fluctuations of the posterior distributions. This suggests that the exploration of TS is controlled automatically, without the need to control significance levels, unlike the UCB approach. In other words, the posterior distribution is dedicated to controlling the dilemma of exploration and exploitation, which highlights its importance in TS. Since the choice of the prior is closely related to the property of the posterior and is a huge topic in the Bayesian approach, we introduce some important priors for this thesis in Section 2.3.

Although the behavior of TS varies with the choice of priors, its core approach remains consistent: just playing an arm according to the given posterior probability distribution. However, computing the posterior probability of an arm being optimal directly is often complex. As a result, an alternative approach, shown in Algorithm 1, is used to implement TS. Firstly, TS generates a Thompson sample  $\tilde{\theta}_i(t)$  from the posterior distribution of the parameter  $\theta_i$ , where  $\nu_{\tilde{\theta}_i(t)}$  represents an estimate of the TS policy for the reward distribution of the arm  $i$  at round  $t$ . Then, TS plays an arm with the largest expected reward, which is formulated by

$$i(t) = \arg \max_{i \in [K]} \mu(\tilde{\theta}_i), \quad \tilde{\theta}_i \sim \pi(\theta_i \mid X_{i,n}),$$

where  $\pi(\theta_i \mid X_{i,n})$  denotes the posterior distribution of parameter  $\theta_i$  after  $n$  observations from the arm  $i$ ,  $X_{i,n} = (x_{i,1}, \dots, x_{i,n})$ . Not only for its simplicity in implementation, but TS has also been widely adopted for its outstanding empirical performance [Chapelle and Li, 2011, Russo et al., 2018].

## 2.3 Thompson Sampling and Priors

Since TS is a Bayesian policy, the choice of the priors is a fundamental problem in the implementation of TS. While one might have their own belief or prior knowledge on parameters that can be utilized to design the prior for the specific problem, such information would not always be available in practice. Therefore, in this section, we introduce some renowned priors, where the primary focus is on *noninformative* priors that do not assume any prior information on the unknown parameters. Then, we review the previous

findings on how the choice of priors affects the performance of TS in different bandit problems. Hereafter, we denote the probability density function of distribution  $\nu_\theta$  by  $f_\theta(\cdot)$ .

### 2.3.1 Conjugate prior

While it is true that prior distributions can be designed arbitrarily, it is obvious that certain choices can make it difficult to accurately infer parameters or induce extremely complicated posterior distributions obtained by Bayes' theorem, which is computationally expensive to handle. The *conjugate prior* is defined to solve at least the latter problem by simplifying the computation of the posterior distribution and performing analytical calculations efficiently. In other words, the conjugate prior is for mathematical convenience, providing the closed-form expression of the posterior distribution. More formally, when  $\pi_c(\theta)$  is a conjugate prior for the likelihood of  $f_\theta(X)$ , then the corresponding posterior  $\pi_c(\theta|X)$  belongs to the distribution family including  $\pi_c(\theta)$ .

For instance, when  $\nu_\theta$  belongs to the Bernoulli distribution where  $\theta \in [0, 1]$  denotes the probability of success, the conjugate prior distribution is a beta distribution  $\text{Beta}(\alpha, \beta)$  with prior hyperparameters  $\alpha, \beta \in \mathbb{R}_+$ . Then, by Bayes' theorem, one can derive the posterior distributions after  $n$  i.i.d. observations,

$$\begin{aligned} \pi(\theta|X_n) &\propto \pi(\theta) \prod_{i=1}^n f_\theta(x_i) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \\ &\propto \text{Beta}\left(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i\right), \end{aligned} \quad (2.4)$$

which shows both the conjugate posterior and prior belong to the Beta distribution. Here, one can combine their belief or knowledge to set prior hyperparameters  $\alpha$  and  $\beta$ , which implies that the conjugate prior is not a noninformative prior [Robert, 2007]. Beyond the algebraic convenience, Agarwal and Daumé [2010] provided a geometric meaning of the conjugate prior, where they showed that the conjugate prior has the same geometry as the likelihood when it belongs to the exponential family.

### 2.3.2 Noninformative prior

However, in the case of non-exponential families, the existence of the conjugate prior is not generally guaranteed. Furthermore, the choice of prior hyperparameters remains an unjustifiable problem without any prior knowledge [Robert, 2007]. Then, the natural question arises: how to design a prior distribution without any prior knowledge of the parameters for general distributions? Since the only available information is the statistical model to which  $f_\theta$  belongs, one should utilize this information in the design of priors. Such priors are called *objective* priors or *noninformative* priors, as they do not take any subjective knowledge into account. Instead, they are usually constructed under certain criteria, where the three most important criteria are simplicity, generality, and trustworthiness [Berger and Bernardo, 1992].

Before introducing renowned approaches, we present a quote describing the interpretation of noninformative priors, which is closely related to the research question addressed in Chapters 3 and 4 of this thesis:

Noninformative priors should be taken as reference<sup>1</sup> or default priors, upon which everyone could fall back when the prior information is missing.

---

<sup>1</sup>The term “reference” here means a standard in the dictionary sense, which is distinct from the reference prior discussed later in this section.

This point was originally noted by Kass and Wasserman [1996] and has been reproduced by Robert [2007, Section 3.5]. In this study, we translate this description to the usefulness of TS with noninformative priors as a *starting point* for bandit problems where no prior knowledge is available.

### Uniform prior

It is known that Laplace [1820] was the first person who used the noninformative prior, which simply assigns equal probability to all possible values within the parameter space. This *uniform prior* obviously represents the ignorance of the parameters and can be defined for any models by  $\pi_u(\theta) \propto 1$ . From its simplicity, the uniform prior has been adopted in various problems such as the inventory modeling [Hill, 1997], the bandit problems [Kaufmann et al., 2012b], and the object recognition [Maturana and Scherer, 2015].

Despite its simplicity and generality, the uniform prior has been criticized due to its variance under reparameterization [Syversveen, 1998]. This means that uniform priors can vary depending on the parameterization of the model. Therefore, when the same model is described by different parameters, the resulting posterior distributions may also be different. The following example illustrates this fundamental problem of the uniform prior.

**Example 1.** *Let us consider the statistical model  $\mathfrak{M}$ . An agent A models  $\mathfrak{M}$  by the parameter  $\theta$  and uses the uniform prior on  $\theta$  parameterization,  $\pi_{u,A}(\theta) = 1$ . On the other hand, another agent B considers the same model  $\mathfrak{M}$  but with different parameterization  $\eta = g(\theta)$ , which is obtained by any one-to-one transformation  $g$ . By the transformation formula, the uniform prior of agent A under agent B’s model can be written as*

$$\pi_{u,A}(\eta) = |\det \nabla g^{-1}(\eta)| \cdot \pi_{u,A}(\theta),$$

where  $\nabla g^{-1}$  denotes the Jacobian matrix of  $g^{-1}$ . Therefore,  $\pi_{u,A}(\eta)$  is usually different from the uniform prior on  $\eta$  parameterization,  $\pi_{u,B}(\eta) = 1$ .

This example provides a clear illustration of the dependence of the uniform prior on the parameterization of the statistical model. Additionally, one must carefully consider which parameterization is appropriate to use with the uniform prior, and it is unclear how to evaluate the appropriateness of parameterizations. As noted in Robert [2007, Section 3.5.1], this issue becomes even more critical when performing inference on multiple parameters.

### Jeffreys prior

To develop an invariant noninformative prior, Jeffreys [1961] utilized the Fisher information (FI) matrix, which does not rely on any prior information about unknown parameters. Note that there are several definitions of the FI, where Schervish [2012] defined the FI matrix as a covariance of the score function under the FI regularity conditions provided in Definition 2.3. However, the definition we adopt in this thesis, as suggested by Lehmann and Casella [2006], is more flexible than some other definitions, and it can accommodate weaker assumptions and constraints. It is worth noting that the FI derived from the following definition is not always well-defined since some elements can be infinite under some distributions.

**Definition 2.2** (Fisher information matrix [Lehmann and Casella, 2006]). For a random variable  $X$  with density  $f_\theta(\cdot)$ , Fisher information that  $X$  contains about the parameter  $\theta$

is defined as

$$[I(\theta)]_{ij} = I_{ij} = \mathbb{E}_\theta \left[ \left( \frac{\partial}{\partial \theta_i} \log f_\theta(X) \right) \left( \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right) \right],$$

where the partial derivative of  $\log f$  denotes the score with respect to parameter  $\theta_i$ . The element  $I_{ij}$  is also called the Fisher information metric in the field of information geometry [Amari, 2016].

Notice that the FI is invariant for all non-singular transformations of the parameters. Here, if  $\eta = g(\theta)$  holds for a differentiable function  $h$ , then the FI that  $X$  contains about  $\eta$  is given as follows [Robert et al., 2009]:

$$\begin{aligned} I(\theta) &= (\nabla g(\theta)) I(\eta) (\nabla g(\theta))^\top \\ \det(I(\theta)) &= \det(I(\eta)) \det(\nabla g(\theta))^2, \end{aligned} \quad (2.5)$$

where  $\nabla g$  denotes the Jacobian matrix of  $g$  and superscript  $\top$  denotes the transpose of a matrix. Therefore, to satisfy the requirement of the invariance on reparameterization, the *Jeffreys prior* is defined as

$$\pi_j(\theta) \propto \sqrt{\det(I(\theta))}.$$

The following example shows the invariance of the Jeffreys prior under any differentiable transformation.

**Example 2 (Continued).** *Instead of the uniform prior, agents A and B use the Jeffreys prior on  $\theta$ ,  $\eta$  parameterizations, respectively, where  $\pi_{j,A}(\theta) \propto \sqrt{\det(I(\theta))}$  and  $\pi_{j,B}(\eta) \propto \sqrt{\det(I(\eta))}$ . Then, the Jeffreys prior of agent A under agent B's model can be written as*

$$\begin{aligned} \pi_{j,A}(\eta) &= |\det(\nabla g^{-1}(\eta))| \cdot \sqrt{\det(I(\theta))} && \text{by transformation formula} \\ &= \sqrt{\det(I(\theta)) (\det(\nabla g^{-1}(\eta))^2)} \\ &= \sqrt{\det(I(\eta))} = \pi_{j,B}(\eta). && \text{by (2.5)} \end{aligned}$$

Therefore, the Jeffreys prior is invariant under any one-to-one transformation.

It is known that the FI matrix defined in (2.2) has alternative expressions that are more convenient to compute if the model satisfies the FI regularity conditions.

**Definition 2.3** (FI regularity conditions [Schervish, 2012]). The following conditions will be known as the FI regularity conditions:

1. There exists  $B$  with  $\nu_\theta(B) = 0$  such that for all  $\theta$ ,  $\frac{\partial f_\theta(x)}{\partial \theta_i}$  exists for  $x \notin B$  and each  $i$ .
2.  $\int f_\theta(x) d\nu_\theta(x)$  can be differentiated under the integral sign with respect to each coordinate of  $\theta$ .
3. The support of  $f_\theta$  is the same for all  $\theta \in \Theta$ .

If the FI regularity conditions hold, it holds for any  $i \in [d]$  that

$$\mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_i} \log f_\theta(X) \right] = 0.$$

Therefore, FI can be written as follows [see Lehmann and Casella, 2006, 6.10.]:

$$I(\theta)_{ij} = \text{cov}_\theta \left[ \frac{\partial}{\partial \theta_i} \log f_\theta(X), \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right].$$

In addition, if  $f_\theta$  is twice differentiable with respect to  $\theta$ , then it coincides with the negative expected value of the Hessian matrix of  $\log f(X|\theta)$ , i.e.,

$$I(\theta) = -\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right].$$

Therefore, if FI regularity conditions hold, one can compute the FI matrix easily and can use it to derive the Cramér–Rao bound on the variance of unbiased estimators [Cramér, 1946, Rao, 1945]. However, if the FI regularity conditions are violated, the FI may not be well-defined or can be singular [see Schervish, 2012, Example 2.81.].

**Example 3.** *Let us consider the model of the uniform distribution  $\text{Uniform}_{a,b}(a, b)$  with density  $f_{a,b}^U(x) = \frac{1}{b-a} \mathbb{1}[a \leq x \leq b]$  for the support  $[a, b]$ , which violates the FI regularity conditions 1 and 3 in Definition 2.3. Here, the FI matrix for each parameterization can be computed from Definition 2.2 as*

$$I(a, b) = \begin{bmatrix} \frac{1}{(b-a)^2} & -\frac{1}{(b-a)^2} \\ -\frac{1}{(b-a)^2} & \frac{1}{(b-a)^2} \end{bmatrix}.$$

*Since the determinant of the FI matrix is zero, one cannot apply the Jeffreys prior directly to the model of the uniform distribution.*

Although the Jeffreys prior is widely applicable, it cannot be generalized to the models that do not satisfy some of FI regularity conditions as shown in Example 3. Furthermore, the Jeffreys prior is known to suffer from many problems when the model contains nuisance parameters [Datta and Ghosh, 1995, Ghosh, 2011]. Here, we provide two famous examples where the usage of the Jeffreys prior induces undesirable results.

**Example 4** (Example 2 in Neyman and Scott [1948]). *Consider the Gaussian model with observations*

$$x_{ij} \sim \text{Gaussian}(\mu_i, \sigma), \quad \forall i \in [n], j \in [2],$$

*where all observations are mutually independent and characterized by a single unknown scale parameter. In this model, sufficient statistics are given as*

$$T((x_{ij})_{(i,j) \in [n] \times [2]}) = (\hat{x}_1, \dots, \hat{x}_n, S^2),$$

*where  $\hat{x}_i = \frac{x_{i1} + x_{i2}}{2}$  and  $S^2 = \sum_{i=1}^n \sum_{j=1}^2 (x_{ij} - \hat{x}_i)^2$ . Here, Example 3 in Berger and Bernardo [1992] showed that when one uses the Jeffreys prior,  $\pi_j(\mu_1, \dots, \mu_n, \sigma) \propto \sigma^{-n+1}$ , the posterior mean for  $\sigma^2$  is given as*

$$\mathbb{E} [\sigma^2 \mid (x_{ij})_{(i,j) \in [n] \times [2]}] = \frac{S^2}{2n-2}.$$

*Since  $\frac{S^2}{n} \rightarrow \sigma^2$  holds as  $n \rightarrow \infty$  with probability one, using the Jeffreys prior leads to inconsistent estimators for the variance in this example, which is known as the Neyman–Scott problem.*



**Example 5** (Stein’s example [Stein, 1956, 1959]). Let  $x_1, \dots, x_n$  be independently normally distributed real random variables with unknown means  $\mu_1, \dots, \mu_n$  and common variance 1, which can be simply denoted by  $x \sim \text{Gaussian}(\mu, I_n)$  for the identity matrix  $I_n$  with rank  $n$ . When the parameter of interest is  $\|\mu\|^2$ , Example 3.5.9 in Robert [2007] showed that the posterior mean of  $\|\mu\|^2$  is  $\|\mu\|^2 + n$ . However, in this problem, the best estimator among the estimators of the form  $\|\mu\|^2 + c$  for the quadratic loss is known as  $\|\mu\|^2 - n$ , which uniformly dominates the generalized Bayesian estimator under the Jeffreys prior.

These challenging examples imply that the Jeffreys prior may not perform well for some inferential problems, particularly when the model contains nuisance parameters. This casts doubt on the reliability of the Jeffreys prior as a guideline prior, despite its desirable properties in regular models without nuisance parameters.

### Reference prior

To develop a method of noninformative priors that can be applied in almost any situation, Bernardo [1979] proposed a *reference prior* approach. Before defining the reference prior, we first introduce the Shannon expected information, which is related to the key idea of the reference analysis. Note that it has several desirable properties such as invariance, non-negativity, and concavity [Bernardo, 1979, see Section 2 and references therein].

**Definition 2.4** (Expected information [Berger et al., 2009]). The expected information to be obtained from one observation from a model  $\mathfrak{M} \equiv \{f_\theta(x) : x \in \mathcal{X}, \theta \in \Theta\}$ , when the prior for  $\theta$  is  $\pi(\theta)$ , is

$$\begin{aligned} \text{EI}(\pi|\mathfrak{M}) &:= \int \int_{\mathcal{X} \times \Theta} f_\theta(x) \pi(\theta) \log \frac{f_\theta(x)}{\pi(\theta)} dx d\theta \\ &= \mathbb{E}_x [\text{KL}(\pi(\theta|x); \pi(\theta))] . \end{aligned}$$

Then, the expected information about  $\theta$  that can be obtained from  $k$  independent observations generated by the model  $\mathfrak{M}$  is denoted by  $\text{EI}(\pi|\mathfrak{M}^k)$ .

Bernardo [1979] interpreted  $\text{EI}(\pi|\mathfrak{M}^k)$  as a measure of the amount of *missing* information about  $\theta$  given prior  $\pi(\theta)$  based on  $k$  independent observations. Then, they suggested using a prior that maximizes  $\lim_{k \rightarrow \infty} \text{EI}(\pi|\mathfrak{M}^k)$  that is missing information after infinite observations. However,  $\lim_{k \rightarrow \infty} \text{EI}(\pi|\mathfrak{M}^k)$  is usually infinite for the continuous parameter space  $\Theta$  since knowing a real number *exactly* requires an infinite amount of information. Instead, Bernardo [1979] defined a *reference posterior* as a limiting result and then defined a reference prior that produces the reference posterior via Bayes’ theorem as follows.

**Definition 2.5** (Reference posterior and prior [Bernardo, 1979]). The reference posterior of  $\theta$  after  $n$  observations is defined as

$$\pi_r(\theta|X_n) = \lim_{k \rightarrow \infty} \pi_{r,k}(\theta|X_n),$$

where  $\pi_{r,k}(\pi|X_n) \propto f_\theta(X_n) \pi_{r,k}(\theta)$  is the posterior density corresponding to that prior  $\pi_{r,k}(\theta)$  which maximizes  $\text{EI}(\pi|\mathfrak{M}^k)$  for  $\pi$  satisfying  $\int_{\Theta} f_\theta(X_n) \pi(\theta) < \infty$ . A reference prior for  $\theta$  is a positive function  $\pi_r(\theta)$  which satisfies  $\pi_r(\theta | X_n) \propto f_\theta(X_n) \pi_r(\theta)$ .

The reference priors coincide with the Jeffreys prior for continuous parameter space without any nuisance parameters, and with the uniform prior for finite parameter space

with sufficient regularity [Kass and Wasserman, 1996]. To handle a model containing nuisance parameters, where the Jeffreys prior suffers from several problems, Bernardo [1979] extended the reference prior by separating parameters into groups. For instance, let us consider the two-parameter model  $\Theta \subseteq \mathbb{R}^2$  where the parameter of interest is  $\theta \in \mathbb{R}$ , and the nuisance parameter is  $\phi \in \mathbb{R}$ . Then, they defined the reference prior by reducing the two-parameter priors to a sequential application of the reference prior for the single-parameter,  $\pi_r(\theta, \phi) = \pi_r(\phi|\theta)\pi_r(\theta)$  [Bernardo and Smith, 2009, see Section 3.8]. To put it simply, one first finds a conditional reference prior for nuisance parameter  $\phi$  with fixed  $\theta$ ,  $\pi_r(\phi | \theta)$ . Based on this conditional prior, one can obtain the marginalized model by

$$f_\theta(x) = \int f_{\theta,\phi}(x)\pi_r(\phi | \theta)d\phi.$$

Again, one can find a reference prior  $\pi_r(\theta)$  for this marginalized model and the reference prior  $\pi_r(\theta, \phi)$  for the original model. This idea can be generalized to the multi-parameter model by separating parameters into groups and applying the same procedure recursively. The general derivations of the reference prior were described by four steps [Berger and Bernardo, 1992], which makes it possible to apply to various models. Furthermore, the reference prior does not suffer from the problems that exist when one uses the Jeffreys prior [Berger and Bernardo, 1992] and is invariant under one-to-one reparameterization if it does not change the ordering of parameters [see Datta and Ghosh, 1996, Theorem 2.1].

Although the derivation of the reference prior is not simple in general, it can be easily computed based on the FI matrix under certain conditions.

**Theorem 2.6** (Theorem 1 of Datta and Ghosh [1995]). *Suppose that the parameters  $\theta = (\theta_1, \dots, \theta_d) \in \Theta$  is group ordered as  $\theta = \{\theta_{(1)}, \dots, \theta_{(m)}\}$ , where  $\theta_{(m)}$  has  $d_i$  coordinates and  $\sum_{i=1}^m d_i = d$ . Here, the subscript  $(i)$  represents a prioritization of inference, where there is greater interest in inference regarding  $\theta_{(i)}$  than in  $\theta_{(i+1)}$ <sup>2</sup>, and where all  $\theta_j$  in the same group have equal importance. Assume that*

- $I(\theta) = \text{diag}(h_1(\theta), \dots, h_m(\theta))$ , where  $h_1(\theta)$  is  $d_1 \times d_1$  matrix than is not necessarily diagonal.
- $\det(h_j(\theta)) = h_{j1}(\theta_{(j)})h_{j2}(\theta_{\sim(j)})$  for nonnegative functions  $h_{j1}$  and  $h_{j2}$  and  $\theta_{\sim(j)} = (\theta_{(1)}, \dots, \theta_{(j-1)}, \theta_{(j+1)}, \theta_{(m)})$ .

Then,

$$\pi_r(\theta) = \prod_{j=1}^m \sqrt{h_{j1}(\theta_{(j)})}. \quad (2.6)$$

**Example 6** (Location-scale family [Ghosh, 2011]). *Suppose that the random variable  $X$  has a probability density function  $\frac{1}{\sigma} f\left(\frac{x-l}{\sigma}\right)$  with location  $l \in \mathbb{R}$  and scale  $\sigma \in \mathbb{R}_+$ . Then, it holds that*

$$I(l, \sigma) = \sigma^{-2} \begin{bmatrix} c_1 & c_2 \\ c_2 & c_3 \end{bmatrix},$$

where  $c_1, c_2$ , and  $c_3$  are functions of  $f$  and do not involve parameters. When  $c_2 = 0$ , from Theorem 2.6, the reference prior is given as  $\pi_r(l, \sigma) \propto \sigma^{-1}$ . Note that even when we changed the order of parameters or  $c_2 \neq 0$ , in the location-scale family cases, it holds that  $\pi_r(\sigma, l) \propto \sigma^{-1}$  [Ghosh, 2011]. On the other hand, the Jeffreys prior is given as  $\pi_j(l, \sigma) = \sqrt{\det I(l, \sigma)} \propto \sigma^{-2}$ .

---

<sup>2</sup>For instance, in the case of the Gaussian distribution with  $\theta = (\mu, \sigma)$ , we can set  $\theta_{(1)} = \mu$  and  $\theta_{(2)} = \sigma$  when our main objective is to estimate  $\mu$ .

## Probability matching prior

The probability matching prior is a type of noninformative prior that is designed to achieve the synthesis between the coverage probability of the Bayesian interval estimates and that of the frequentist interval estimates [Tibshirani, 1989, Welch and Peers, 1963]. Therefore, the posterior probability of certain intervals matches exactly or asymptotically the frequentist's coverage probability under the probability matching prior.

Although several matching priors have been developed under slightly different considerations [see Datta and Mukerjee, 2004, for more details about other variants], we introduce the quantile matching prior, which is a common approach [Ghosh, 2011, Robert, 2007]. The quantile matching prior aims to achieve a synthesis between the credible interval and confidence interval. For any priors  $\pi$ , suppose that

$$\mathbb{P}_{\theta \sim \pi}[\theta \in C_\alpha \mid X] = \int \mathbb{P}[\theta \in C_\alpha] \pi(\theta) d\theta = 1 - \alpha$$

for  $\alpha \in (0, 1)$  and a set  $C_\alpha \subset \Theta$ . When the prior is a probability matching prior,  $\pi_{\text{pm}}$ , it holds that

$$\mathbb{P}[\theta \in C_\alpha] = \alpha + \mathcal{O}(n^{-(k+1)/2}). \quad (2.7)$$

A prior satisfying (2.7) is called the  $k$ -th order matching prior<sup>3</sup>. Any positive continuous prior  $\pi$  satisfies the zeroth order matching property from the first-order quadratic approximation, which shows the equivalence of frequentist and Bayesian normal approximation up to  $\mathcal{O}(n^{-1/2})$ . Furthermore, the first-order matching prior is known to be invariant under one-to-one reparameterization [Datta and Ghosh, 1996]. If there are no additional terms in (2.7), such a prior is called an exact matching prior.

When there are no nuisance parameters, the Jeffreys prior is known to be the unique first-order matching prior [Datta and Mukerjee, 2004]. In the presence of the nuisance parameters, Peers [1965] showed that the first-order matching prior is equivalent to the solution of a partial differential equation. Therefore, it is usually difficult to derive the probability matching priors, which becomes more complex when one considers the multiparameter models. Nevertheless, when the FI matrix is diagonal, the unique first-order joint probability matching prior is given as follows [Datta and Sweeting, 2005]:

$$\pi_{\text{jpm}}(\theta) \propto \prod_{j=1}^d \sqrt{h_{j1}(\theta_j)},$$

which is the same as the reference prior given in (2.6) when every parameter group is a singleton. Furthermore, when  $\theta_1$  is a parameter of interest and  $\theta_{\sim(1)} = (\theta_2, \dots, \theta_d)$  be a vector of nuisance parameters, the first-order probability matching prior is given as follows [Nicolaou, 1993, Tibshirani, 1989]:

$$\pi_{\text{pm}}(\theta) = g(\theta_{\sim(1)}) \sqrt{h_{11}(\theta)}, \quad (2.8)$$

where  $g(\cdot)$  is an arbitrary positive function. Note that, in the absence of orthogonalization,  $g(\cdot)$  cannot be arbitrary in general [Datta and Sweeting, 2005]. However, for the location-scale family considered in Example 6, it is known that the unique second-order probability matching prior is given as  $\pi_{\text{pm}}(l, s) \propto \sigma^{-1}$  regardless of orthogonality [Datta and Mukerjee, 2004]. Furthermore, DiCiccio et al. [2017] showed that this prior yields exact conditional matching in the univariate location-scale family regardless of parameters of interest.

---

<sup>3</sup>Note that some papers call a prior the  $k$ -th order matching prior when a remainder is  $\mathcal{O}(n^{-k/2})$  [Datta and Sweeting, 2005]. Here, we follow the notations used in DiCiccio et al. [2017], Ghosh [2011], and Mukerjee and Ghosh [1997]

Table 2.1: Asymptotic optimality of Thompson sampling with different noninformative priors for various models. ✓ and ✗ denote whether TS can achieve the asymptotic lower regret bound in (2.1) for the corresponding rewards model or not.

Model	Parameter $\theta$	Priors	Optimal
Bernoulli [Kaufmann et al., 2012b]	prob. of success $p \in [0, 1]$	uniform prior	✓
SPEF [Korda et al., 2013]	$\theta \in \mathbb{R}$	Jeffreys prior	✓
Multinomial	prob. vector $p \in \Sigma_M$ for given $M \in \mathbb{N}$	uniform prior	✓
Bounded support [Riou and Honda, 2020]	nonparametric known support $[0, B]$	uniform prior	✓
Gaussian [Honda and Takemura, 2014]	location (mean) and scale $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$	uniform prior Jeffreys prior reference prior	✓ ✗ ✗

### 2.3.3 Performance of Thompson sampling

Since TS is a Bayesian randomized policy, the Bayesian regret of TS has been considered in several bandit models such as Gaussian distributions with unit variance [Bubeck and Liu, 2013], linear models [Russo and Van Roy, 2014], and symmetric  $\alpha$ -stable distributions [Dubey and Pentland, 2019]. However, as discussed in Section 2.1.1, the Bayesian regret of a policy can be less informative in the sense that a bound on the Bayesian regret could provide a loose bound on the problem-dependent regret in general. Therefore, the performance of TS has been analyzed from the frequentist’s perspective, in terms of problem-dependent regret.

The asymptotic optimality of TS has been shown in several reward models such as Bernoulli distributions [Kaufmann et al., 2012b], single parameter canonical exponential families [Korda et al., 2013], and multinomial distributions [Riou and Honda, 2020]. Most of the previous research focused on the analysis of TS based on the conjugate prior with specific hyperparameters since TS with those choices of priors was sufficient to show the asymptotic optimality in the single parameter models. For example, Kaufmann et al. [2012b] considered the uniform prior that is equivalent to a conjugate prior of the Bernoulli distribution, Beta distribution, with hyperparameters  $(\alpha, \beta) = (0, 0)$  in (2.4). The asymptotic optimality of TS was extended to the single parameter exponential family (SPEF), with the choice of the Jeffreys prior [Korda et al., 2013]. Since the Bernoulli distribution belongs to the SPEF, such results would imply that the performance of TS is not that sensitive to the choice of noninformative priors unless one selects a well-studied noninformative prior.

However, Honda and Takemura [2014] showed that such observations do not hold in the Gaussian models with unknown mean and variance, which is a two-parameter exponential distribution belonging to the location-scale family. To be precise, they showed the suboptimality of the Jeffreys prior and the reference prior, which was optimal in the SPEF. These results invoke the importance of understanding how the noninformative affects the performance of TS in bandit problems, as well as that of how to choose a noninformative prior when we face a new model. The currently known optimal results of TS are summarized in Table 2.1. Although TS with the uniform prior seems a reasonable choice in the bandit problems, the usage of the uniform prior essentially associates with the way of parameterization, which has a possibility to incur some problems for other models. In the following chapters, we discuss the default choice of noninformative priors for TS in MAB problems, which can serve as a fallback option when dealing with *new bandit models*.

## Chapter 3

### Uniform Bandits

In this chapter, we study the asymptotic optimality of Thompson sampling (TS) for the bandit model of the uniform distribution with unknown supports. To the best of this author’s knowledge, this is the first study that examines TS on non-regular models that do not satisfy the Fisher information regularity condition in Definition 2.3. Our finding also shows that the optimality of TS depends on the choice of noninformative priors, where the uniform prior is optimal while the reference prior and the Jeffreys prior are suboptimal. These results coincide with the previous findings in the Gaussian bandits [Honda and Takemura, 2014].

However, as discussed in Section 2.3.2, the uniform prior is inherently dependent on the specific parameterization of the distributions. While TS with the uniform prior with location-scale parameterization shows the optimal results, it does not necessarily guarantee the optimality on different parameterizations. We formulate the suboptimality of the uniform prior with different parameterizations even for the location-rate parameterizations. In light of this limitation, we propose a slightly modified TS-based policy, called TS with Truncation (TS-T), which can achieve the asymptotic optimality for the Gaussian distributions and the uniform distributions by using the reference prior and the Jeffreys prior that are invariant under one-to-one reparameterizations. The pre-processing on the posterior distribution is the key to TS-T, where we add an adaptive truncation procedure on the parameter space of the posterior distributions.

#### 3.1 Introduction

In this section, we provide a brief introduction to the background and motivation behind this chapter, as well as a summary of its key contributions.

##### 3.1.1 Chapter background

The classical parametric stochastic multi-armed bandit (MAB) problems assume that each arm can be specified by distribution  $\nu_\theta$  whose univariate density function is given as  $f_\theta(x)$  for the unknown parameters  $\theta \in \mathbb{R}^d$ . In practice, a specific distribution is used to model the underlying distributions of arms based on their problem. For example, one might consider the Bernoulli distribution with the probability mass function  $f_\theta(x) = \mathbb{1}[x = 1]\theta + \mathbb{1}[x = 0](1 - \theta)$  for  $\theta \in [0, 1]$ . With the assumption on  $f$ , one can choose a specific prior distribution of TS based on their belief or previous experience.

Since prior knowledge on parameters is not always available in practice, this thesis focuses on noninformative priors, which do not assume any information on parameters. Nevertheless, when it comes to the regret bounds of TS, it is often reported that TS is not too sensitive to the choice of the prior for the single parameter models. For example, both the uniform prior [Kaufmann et al., 2012b] and the Jeffreys prior [Korda et al.,

2013] are found to be optimal for models of the Bernoulli distributions. Note that the reference prior also leads to the optimal regret bound for the Bernoulli bandit models since the Jeffreys prior coincides with the reference prior for the one-parameter models. However, it has been shown that the choice of noninformative priors can significantly impact the performance of TS for the Gaussian models [Honda and Takemura, 2014], which is a noncompact multiparameter model. This result indicates that the choice of noninformative priors becomes more challenging in multiparameter models than that in one-parameter models. Nevertheless, to the best of this author’s knowledge, the Gaussian bandit is the only multiparameter bandit model where the problem-dependent optimality of TS has been studied despite its practical usefulness.

### 3.1.2 Chapter contribution

In this chapter, we first show that the prior sensitivity of TS occurs not only in the non-compact multiparameter models but also in the uniform model with unknown supports, which is a non-regular compact multiparameter model. Specifically, we show that TS with the uniform prior on the location-scale (LS) parameterization is asymptotically optimal, while TS with the reference prior and the Jeffreys prior are found to be suboptimal. The implication of this discovery is twofold. In the first place, the bounds show the importance of choosing priors in multiparameter models, extending the understanding of TS to non-regular compact multiparameter models. Moreover, the lack of invariance makes the optimal result of the uniform priors less informative, as some uniform priors with different parameterizations result in suboptimal performance. Therefore, for regret minimization in multiparameter models, one might look for a better invariant prior or replace the TS algorithm to obtain a practical and optimal arm selection policy.

However, as mentioned above, well-known invariant priors do not always result in optimal performance under multiparameter bandit models. To address such challenges, we propose a variant of TS, called TS with Truncation (TS-T), for the uniform models and the Gaussian models. We provide a finite-time regret analysis of TS-T, which demonstrates its asymptotic optimality under the reference prior and the Jeffreys prior for both models. Our approach builds upon the basic strategy of TS, but with key modifications that improve the performance and address the limitations of TS. In particular, we devise an adaptive truncation procedure on the parameter space of the posterior distribution to control the problems in the early stage of learning, hence the name truncation in TS-T.

The contributions of this chapter are summarized as follows:

- We prove the asymptotic optimality/suboptimality of TS under different choices of noninformative priors for models of uniform distributions with unknown supports. This extends the understanding of TS in the multiparameter models, which has not been well studied in the literature so far, showing the importance of choosing noninformative priors.
- We explicitly show that some uniform priors with different parameterizations are suboptimal. This makes the optimality of TS with the uniform prior with location-scale (LS) parameterization less attractive in practice when one considers reparameterization for efficient computations.
- We propose a variant of TS that is asymptotically optimal for the uniform models and the Gaussian models under the reference prior and the Jeffreys prior, where the vanilla TS is found to be suboptimal. This provides optimal results that remain consistent regardless of the way of parameterizing the LS family, which addresses the limitations of the vanilla TS.

### 3.1.3 Chapter organization

The rest of this chapter is organized as follows. We formulate the MAB problem for the LS family and introduce the asymptotic optimality for the uniform bandits in Section 3.2. Section 3.3 discusses the choice of priors in the LS family that includes the uniform distributions and the Gaussian distributions as a special case. Then, we propose a variant of TS for the general LS family and instantiate it to consider the uniform bandits. We show the asymptotic optimality of TS and TS-T and the failure of some uniform priors in Section 3.4. Based on observations that our findings under the uniform bandits coincide with the behavior of TS under the Gaussian bandits, we further show the asymptotic optimality of TS-T with the reference prior and the Jeffreys prior under the Gaussian bandits in Section 3.5. Numerical results in Section 3.6 support our theoretical analysis, where TS-T performs well even in the small time horizon. All the detailed proofs in this chapter are provided in Section 3.7.

## 3.2 Problem Formulation

In this section, we formulate  $K$ -armed bandit problems and the asymptotic regret lower bound for the model of distributions belonging to the LS family for later discussions. Then, we instantiate it to the uniform bandits.

Suppose that there are finite  $K$  arms associated with a distribution belonging to the LS family, where each arm can be represented by a density function  $f_{l,\sigma}(x)$  with location  $l \in \mathbb{R}$  and scale  $\sigma \in \mathbb{R}_+$ . Here, the parameters  $(l, \sigma) \in \mathbb{R} \times \mathbb{R}_+$  are unknown to the agent. Note that we consider MAB problems where every arm is modeled by the *same* type of distribution but with possibly different parameters. Therefore, if the reward distribution of an arm follows the Cauchy distribution, the corresponding distributions of other  $K - 1$  arms also follow the Cauchy distribution.

If a random variable  $X$  with the density function  $f_\theta(x)$  belongs to the LS family, then  $f_{l,\sigma}$  can be written using a probability density function  $f_{0,1}(\cdot)$  as

$$f_{l,\sigma}(x) = \frac{1}{\sigma} f_{0,1}\left(\frac{x-l}{\sigma}\right). \quad (3.1)$$

Note that location  $l$  is not necessarily equivalent to the expectation  $\mu(\theta) = \mathbb{E}_{\nu_\theta}[X]$ . For example, in the Cauchy distribution, the expectation  $\mu$  is undefined, but the location parameter is still finite. One can retrieve the density function of the Gaussian distribution  $\text{Gaussian}(\mu, \sigma)$  with location (mean)  $\mu$  and scale  $\sigma$ ,  $f_{\mu,\sigma}^G(x)$ , by substituting the standard normal density for  $f_{0,1}$ . The uniform distribution can be obtained by letting  $f_{0,1}(x) = \mathbb{1}[0 \leq x \leq 1]$ . If  $X$  follows the uniform distribution  $\text{Uniform}_{\mu\sigma}(\mu, \sigma)$  under the LS parameterization, then it has the density of the form with location (mean)  $\mu$  and scale  $\sigma$ ,

$$f_{\mu,\sigma}^{\text{U}}(x) = \frac{1}{\sigma} \mathbb{1}\left[\mu - \frac{\sigma}{2} \leq x \leq \mu + \frac{\sigma}{2}\right].$$

The uniform distribution can be reparameterized in terms of the boundary of the support by letting  $(a, b) = (\mu - \frac{\sigma}{2}, \mu + \frac{\sigma}{2})$ , denoted by  $\text{Uniform}_{ab}(a, b)$ , whose density function is given as

$$f_{a,b}^{\text{U}}(x) = \frac{1}{b-a} \mathbb{1}[a \leq x \leq b].$$

Here, we assume that the arm 1 has the maximum expected reward for convenience without loss of generality, i.e.,  $\mu_1 = \max_{i \in [K]} \mu_i$ . This assumption is for the simplicity of the analysis, and it is worth noting that incorporating additional optimal arms can only decrease the expected regret of TS [see Agrawal and Goyal, 2012, Appendix A]. By explicitly computing the infimum over the KL divergence, the problem-dependent

regret lower bound in (2.1) can be written in the closed form under uniform models as follows [Cowan and Katehakis, 2015]:

$$\begin{aligned} \liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} &\geq \sum_{i=2}^K \frac{\Delta_i}{\inf_{(\mu, \sigma): \mu > \mu_1} \text{KL}(\text{Uniform}_{\mu\sigma}(\mu_i, \sigma_i); \text{Uniform}_{\mu\sigma}(\mu, \sigma))}, \\ &= \sum_{i=2}^K \frac{\Delta_i}{\log \left(1 + \frac{2\Delta_i}{\sigma_i}\right)}. \end{aligned} \quad (3.2)$$

Recall that an algorithm is called asymptotically optimal over the uniform models if it satisfies

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} = \sum_{i=2}^K \frac{\Delta_i}{\log \left(1 + \frac{2\Delta_i}{\sigma_i}\right)}.$$

### 3.3 Thompson Sampling and the Choice of Priors

In this section, we instantiate TS and propose a variant of TS, TS-T, for the uniform model with the choice of noninformative priors.

As TS is a policy derived from the Bayesian perspective, the choice of the prior is crucial. Since prior knowledge on parameters is not always available in practice, this thesis focuses on noninformative priors, which do not assume any information on parameters. Although the uniform distribution does not satisfy the FI regularity condition, we can derive the FI matrix based on Definition 2.2, which is given as

$$I(\mu, \sigma) = \begin{bmatrix} 0 & 0 \\ 0 & \sigma^{-2} \end{bmatrix}.$$

Note that one can obtain the same results by letting  $c_1 = c_2 = 0$  in Example 6, which provides the FI matrix for the LS family.

Based on this FI matrix, one can derive the reference prior  $\pi_r \propto \sigma^{-1}$ , which coincides with the first-order probability matching prior in (2.8). Although the Jeffreys prior is not well-defined for the uniform distribution, we use the Jeffreys prior  $\pi_j \propto \sigma^{-2}$ , which is the Jeffreys prior for the regular LS family. Unless otherwise stated,  $\pi_u \propto 1$  denotes the uniform prior with  $(\mu, \sigma)$  parameterization. Therefore, we choose the priors of the form  $\sigma^{-k}$  for finite  $k \in \mathbb{R}$ , which includes the renowned noninformative priors as a special case. Under the priors  $\sigma^{-k}$ , the joint posterior after observing  $n$  rewards from the arm  $i$ ,  $X_{i,n} := (x_{i,1}, \dots, x_{i,n})$  is given as

$$\pi^k(\mu, \sigma | X_{i,n}) := \frac{\frac{1}{\sigma^k} \prod_{s=1}^n f_{\mu, \sigma_i}(x_{i,s})}{\iint \frac{1}{\sigma^k} \prod_{s=1}^n f_{\mu, \sigma}(x_{i,s}) d\mu d\sigma} = \frac{\frac{1}{\sigma^{n+k}} \prod_{s=1}^n f_{0,1}\left(\frac{x_{i,s} - \mu_i}{\sigma_i}\right)}{\iint \frac{1}{\sigma^{n+k}} \prod_{s=1}^n f_{0,1}\left(\frac{x_{i,s} - \mu_i}{\sigma_i}\right) d\mu d\sigma},$$

where we denoted the density function of any distribution in the LS family by  $f_{\mu, \sigma}(\cdot)$  and used the property of the LS family in (3.1). Let us denote the (classical) sufficient statistic  $T(X_{i,n})$  for the parameter  $(\mu_i, \sigma_i)$ , which is Bayes-sufficient [Blackwell and Ramamoorthi, 1982]. Then, we can rewrite the posterior distribution using the sufficient statistic as

$$\pi_{i,n}(\mu, \sigma) := \pi(\mu, \sigma | X_{i,n}) = \pi(\mu, \sigma | T(X_{i,n})).$$

Recall Algorithm 1, which illustrates the basic sampling rule of TS where the vanilla TS observes samples  $(\tilde{\mu}_i(t), \tilde{\sigma}_i(t))$  generated from the posterior distribution  $\pi_{i,N_i(t)}(\mu, \sigma)$  at every round  $t$ . However, due to its parametric support, which violates the regularity condition, MLEs are not determined uniquely [Schervish, 2012]. Nevertheless, if any



---

**Algorithm 2** Thompson Sampling for the uniform models

---

- 1: **Parameters:** Set prior  $\pi(\mu, \sigma) = \sigma^{-k}$  for given  $k$ .
  - 2: **Initialization:** Play every arm  $n_0 = \max(2, 3 - \lceil k \rceil)$  times and compute estimators.
  - 3: **for**  $t = n_0 K + 1, \dots, T$  **do**
  - 4: Sample  $\tilde{\sigma}_i(t) \sim \pi^k(\sigma | \hat{\mu}_{i,n}, \hat{\sigma}_{i,n})$  in (3.3) by using the inverse transform sampling.
  - 5: Sample  $\tilde{\mu}_i(t) \sim \text{Uniform}_{\mu\sigma}(\hat{\mu}_{i,n}, \tilde{\sigma} - \hat{\sigma}_{i,n})$
  - 6: Play  $i(t) = \arg \max_{i \in [K]} \tilde{\mu}_i(t)$  and observe a reward  $x_{i(t), N_{i(t)}(t)+1}$ .
  - 7: Update estimators  $\hat{\mu}_{i(t)}$  and  $\hat{\sigma}_{i(t)}$ .
- 

MLE exists, an MLE can be chosen as a function of the sufficient statistic [Moore, 1971]. Therefore, we select MLEs  $(\hat{\mu}_{i,n}, \hat{\sigma}_{i,n})$  that can be expressed by  $T(X_{i,n})$  for the uniform distributions. By abuse of notation, we denote the posterior distribution after  $n$  observations by  $\pi_{i,n}^k = \pi^k(\mu, \sigma | \hat{\mu}_{i,n}, \hat{\sigma}_{i,n})$ , instead of  $\pi(\mu, \sigma | T(X_{i,n}), \hat{\mu}_{i,n}, \hat{\sigma}_{i,n}) = \pi(\mu, \sigma | T(X_{i,n}))$  to explicitly express the estimators after  $n$  observations under the priors  $\sigma^{-k}$ . We use this notation since it makes it easy to see the difference between the vanilla TS and TS-T in the later discussion.

### 3.3.1 Thompson sampling for the uniform bandits

When rewards  $(x_{i,s})$  are generated from  $\text{Uniform}_{\mu\sigma}(\mu_i, \sigma_i)$ , the sufficient statistic is given as  $T(X_{i,n}) = (x_i^{(1)}, x_i^{(n)})$  where  $x_i^{(1)} = \min_{s \in [n]} x_{i,s}$  and  $x_i^{(n)} = \max_{s \in [n]} x_{i,s}$  denote the order statistics. Then, the marginal posterior of  $\sigma$  and the conditional posterior of  $\mu$  given  $\sigma$  under the prior  $\sigma^{-k}$  can be derived as follows:

$$\pi^k(\sigma | \hat{\mu}_{i,n}, \hat{\sigma}_{i,n}) = (n+k-1)(n+k-2) (\hat{\sigma}_{i,n})^{n+k-2} \frac{\sigma - \hat{\sigma}_{i,n}}{\sigma^{n+k}} \mathbb{1}[\sigma \geq \hat{\sigma}_{i,n}], \quad (3.3)$$

$$\pi^k(\mu | \hat{\mu}_{i,n}, \hat{\sigma}_{i,n}, \sigma = \tilde{\sigma}) = f_{\hat{\mu}_{i,n}, \tilde{\sigma} - \hat{\sigma}_{i,n}}^{\text{U}\mu\sigma}(\mu), \quad (3.4)$$

where we used MLEs  $\hat{\mu}_{i,n} = (x_i^{(n)} + x_i^{(1)})/2$  and  $\hat{\sigma}_{i,n} = x_i^{(n)} - x_i^{(1)}$  to be the functions of  $T(X_{i,n})$ . Note that the conditional posterior of  $\mu$  is given as  $\text{Uniform}_{\mu\sigma}(\hat{\mu}_{i,n}, \tilde{\sigma} - \hat{\sigma}_{i,n})$ . However, the marginal posterior of  $\mu$  cannot be represented by a well-known probability distribution, which makes one use Markov chain Monte Carlo methods for calculating its approximations. However, it would make the implementation of algorithms complex and incur high computation costs.

In this chapter, we apply the sequential sampling scheme rather than using computationally costly numerical approximations. This means that  $\tilde{\sigma}$  is sampled first from the marginal posterior in (3.3), which can be easily implemented by using the inverse transform sampling method. Then we sample  $\tilde{\mu}$  from the conditional posterior given the sampled scale parameter in (3.4). This will give the same result when one samples  $\mu$  from the joint posterior by  $\pi_{i,n}(\mu, \sigma) = \pi_{i,n}(\sigma)\pi_{i,n}(\mu|\sigma)$  for  $n \geq n_0 = \max(2, 3 - \lceil k \rceil)$  to avoid improper posteriors, where  $\lceil \cdot \rceil$  denotes the ceiling function. The TS policy for the uniform models is illustrated in Algorithm 2.

### 3.3.2 Thompson sampling with truncation for the LS family

As shown in previous studies [Honda and Takemura, 2014], TS sometimes plays only suboptimal arms when the posterior of the optimal arm has a very small variance in the early stage of learning, which contributes to the suboptimality in expectation. To avoid such problems, we propose a variant of TS, called TS with Truncation (TS-T), where the adaptive truncation is applied to the parameter space of the posterior.

---

**Algorithm 3** Thompson Sampling with Truncation (TS-T) for the LS family

---

- 1: **Parameters:** Set prior  $\pi(\mu, \sigma) = \sigma^{-k}$  for given  $k$ .
  - 2: **Initialization:** Play every arm  $n_0$  times and compute estimators.
  - 3: **for**  $t = n_0K + 1, \dots, T$  **do**
  - 4: Sample  $(\tilde{\mu}_i(t), \tilde{\sigma}_i(t)) \sim \bar{\pi}_{i, N_i(t)}(\mu, \sigma)$  in (3.5).
  - 5: Play  $i(t) = \arg \max_{i \in [K]} \tilde{\mu}_i(t)$  and observe a reward  $x_{i(t), N_{i(t)}(t)+1}$ .
  - 6: Update estimators  $\hat{\mu}_{i(t)}$  and  $\hat{\sigma}_{i(t)}$ .
- 

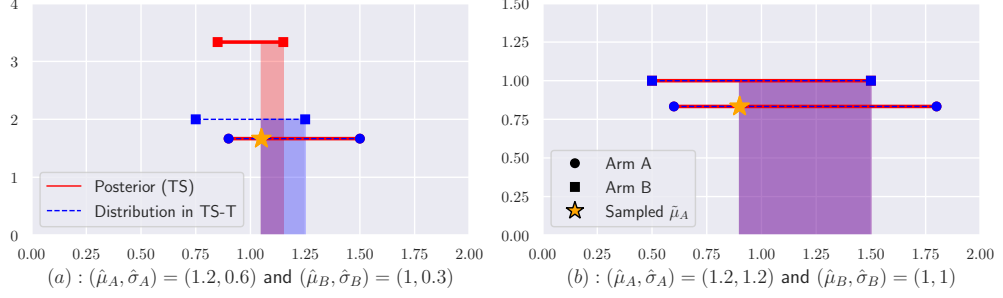


Figure 3.1: A two-armed example where the posterior of  $\mu$  is given as  $\text{Uniform}_{\mu\sigma}(\hat{\mu}, \hat{\sigma})$  and TS-T replaces  $\hat{\sigma}$  with  $\bar{\sigma} = \max(\hat{\sigma}, 0.5)$ . Suppose that  $\hat{\mu}_B = 1 < \hat{\mu}_A = 1.2$  holds in some rounds. The shaded regions denote the probability of the currently suboptimal arm B being played when (a)  $\tilde{\mu}_A = 1.05$  and (b)  $\tilde{\mu}_A = 0.9$  are denoted by a star mark. Different line widths are used to distinguish two lines when they are overlapped.

For the LS family, it can be implemented by sampling parameters from the distributions obtained by replacing an MLE of the scale parameter  $\hat{\sigma}_{i,n}$  with a truncated estimator  $\bar{\sigma}_{i,n}$  satisfying  $\bar{\sigma}_{i,n} = \Omega(n^{-\beta})$  for some  $\beta > 0$ . We choose a specific  $\beta$  to simplify regret analysis, but our discussion can be easily extended to any  $\beta > 0$ . Such truncation prevents an extreme case where  $\hat{\sigma}_{i,n} \approx 0$  for small  $n$ , where the posterior distribution is concentrated on the current mean estimate. In summary, TS-T under the LS family is a policy that samples parameters from the distribution at every round, which is

$$\bar{\pi}_{i,n}(\mu, \sigma) = \pi(\mu, \sigma | \hat{\mu}_{i,n}, \bar{\sigma}_{i,n}). \quad (3.5)$$

Strictly speaking, TS-T is not a Bayesian policy, but a kind of randomized probability matching policy as the corresponding distribution in (3.5) is not a posterior distribution anymore. However, by the choice of  $\bar{\sigma}_{i,n} = \Omega(n^{-\beta})$ , it holds that  $\lim_{n \rightarrow \infty} \bar{\sigma}_{i,n} \geq 0$ , which implies

$$\bar{\pi}_{i,n} \xrightarrow{n \rightarrow \infty} \pi_{i,n}.$$

Therefore, TS-T will behave like TS as  $n$  increases where the truncation has almost no effect. The general TS-T policy for the LS family is illustrated in Algorithm 3, where we instantiate it to the specific distributions in the rest of this section.

The difference from the vanilla TS is that TS-T samples parameters from a distribution obtained by truncating the *parameter space* of the posterior distribution. To be precise, TS directly generates samples from the posterior distributions while TS-T utilizes a modified distribution where the scale parameter space is restricted as  $\mathbb{R}_{\geq n^{-\beta}}$  instead of  $\mathbb{R}_+$ . Therefore, TS-T can be seen as a pre-processed posterior sampling method since we modify the distribution over parameters before sampling. To illustrate the difference, we consider a two-armed example in Figure 3.1, where the probability of the currently suboptimal arm being played is proportional to the area of a shaded region. Figure 1(a) shows that TS-T reserves a larger probability to sample the currently suboptimal arm

than TS, which makes the policy conservative as it avoids insufficient sampling in the early stage of learning. Figure 1(b) illustrates the case where TS-T coincides with TS, which would be the case when both arms are sampled sufficiently, where the truncation has almost no effect. Note that the moderate conservativeness of the policy was found to be a key to the optimality in the Gaussian models [Honda and Takemura, 2014].

**Remark 1. Comparison with a post-processed TS policy** As we explained above, the parameter-sampling distribution  $\bar{\pi}$  considered in TS-T is neither a truncated nor a clipped posterior distribution, which restricts the *support* of the posterior distribution. Jin et al. [2021] considered a TS-based policy, called minimax optimal TS (MOTS), with the clipped posterior distribution, where they reassigned the probability outside of the restricted support to the closest endpoint. The difference between the distributions in TS-based policies is illustrated by Example 7. We select distributions with a truncated parameter space rather than a clipped/truncated posterior since the former can design the conservativeness of the policy by manipulating the flatness of the distributions.

**Example 7.** Assume the posterior distribution of  $\mu$  is given as  $\mathcal{N}(1, \hat{\sigma})$ . Let us consider a truncated estimator  $\bar{\sigma} = \max(\hat{\sigma}, 2)$  for TS-T and restrict the support as  $(\infty, 2]$  for MOTS. Then, when  $\hat{\sigma} = 3$ , TS-based policies generated  $\tilde{\mu}$  from the distributions whose density functions are given as

$$f(\tilde{\mu}) = \begin{cases} f_{1,3}^G(\tilde{\mu}) & \text{under TS and TS-T,} \\ f_{1,3}^G(\tilde{\mu}) \mathbb{1}[\mu \leq 2] + \delta(\mu - 2) \int_2^\infty f_{1,3}^G(\mu) dx & \text{under MOTS,} \end{cases}$$

and when  $\hat{\sigma} = 1$ ,

$$f(\tilde{\mu}) = \begin{cases} f_{1,1}^G(\tilde{\mu}) & \text{under TS,} \\ f_{1,2}^G(\tilde{\mu}) & \text{under TS-T,} \\ f_{1,1}^G(\tilde{\mu}) \mathbb{1}[\mu \leq 2] + \delta(\mu - 2) \int_2^\infty f_{1,1}^G(\mu) dx & \text{under MOTS,} \end{cases}$$

where  $\delta(x)$  denotes the Dirac delta function at  $x$ . TS directly generates parameters from the posterior distribution, while TS-T utilizes a modified distribution where the scale parameter space is restricted as  $\mathbb{R} \geq 2$  instead of  $\mathbb{R}_+$ . Therefore, TS-T can be seen as a pre-processed posterior sampling method. On the other hand, MOTS is equivalent to clipping the output with threshold 1.5 after generating samples from the posterior distribution, which can be seen as a post-processed TS policy.

It is worth noting that MOTS also applied similar pre-processing before clipping, where they added an inflation parameter to enlarge the scale of the distributions that sample parameters. Although the intuition behind is the same as TS-T, it is important to clarify that their approach is specifically designed for the Gaussian model with the uniform prior and known scale  $\sigma = 1$ . This limitation restricts the generalizability of their method to other models, whereas our approach allows for straightforward generalization to other models once posterior distributions are obtained. Note that, simply speaking, our approach can be implemented by replacing an MLE with a truncated estimator, where both can be derived using sufficient statistics.

### 3.4 Main Theoretical Results

This section provides a finite-time regret analysis of TS that shows the asymptotic optimality under the prior  $\sigma^{-k}$  with  $k < 1$  for the uniform models. Then, we show the same results do not hold under  $k \geq 1$ , including the reference prior ( $k = 1$ ) and the Jeffreys prior ( $k = 2$ ). The detailed proofs of theorems in this section are postponed to Section 3.7.

**Theorem 3.1.** Assume that the arm 1 is the unique optimal arm with a finite mean. Given arbitrary  $\epsilon \in \left(0, \min_{i \neq 1} \frac{\Delta_i}{2}\right)$ , the expected regret of TS in Algorithm 2 with the prior  $\sigma^{-k}$  with  $k < 1$  for the uniform models is bounded as

$$\mathbb{E}[\text{Reg}(T)] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{\log \left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)} + \mathcal{O}(\epsilon^{-2}).$$

By letting  $\epsilon = o(1)$  in Theorem 3.1, we see that TS satisfies

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \leq \sum_{i=2}^K \frac{\Delta_i}{\log \left(1 + \frac{2\Delta_i}{\sigma_i}\right)},$$

which shows the asymptotic optimality of TS with  $k < 1$  in terms of the regret lower bound in (3.2). One can obtain an  $\epsilon$ -free version of the regret bound by letting  $\epsilon := \mathcal{O}((\log T)^{-1/3})$  that is tighter than the optimal upper-confidence bound (UCB) based policy of Cowan and Katehakis [2015], where the remainder term is  $\mathcal{O}(\epsilon^{-3})$ .

Next, we show that the vanilla TS with  $k \geq 1$  based on the posteriors in (3.3) and (3.4) cannot achieve the regret lower bound in the theorem below. To simplify the analysis, we consider two-armed bandit problems where two arms have the same left-boundary point of the support. Furthermore, we provide the full information on the arm 2 to the agent, where the prior on the arm 2 is the Dirac measure so that  $\tilde{\mu}_2(t) = \mu_2$  holds for any round  $t \in \mathbb{N}$ .

**Theorem 3.2.** Assume that the arm 1 follows  $\text{Uniform}_{ab}(a_1, b_1)$  and the arm 2 follows  $\text{Uniform}_{ab}(a_2, b_2)$  with  $a_1 = a_2$  and  $b_2 < b_1$ , where  $\mu_1 > \mu_2$  holds. When  $\tilde{\sigma}_1(t)$  and  $\tilde{\mu}_1(t)$  are sampled from the posteriors in (3.3) and (3.4) with the priors  $k \geq 1$ , and  $\tilde{\mu}_2(t) = \mu_2$  holds, there exists a constant  $\xi^U > 0$  independent of  $\sigma_2$  satisfying

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \geq \Delta_2 \xi^U.$$

If  $k > 1$ , then there exist constants  $\xi_k^U > 0$  independent of  $\sigma_2$  satisfying

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{T^{\frac{k-1}{k}}} \geq \Delta_2 \xi_k^U.$$

Theorem 3.2 shows that TS with  $k \geq 1$  suffers at least logarithmic regrets in expectation. Although the regret lower bound in (3.2) approaches zero for sufficiently small  $\sigma_2 = b_2 - a_2$ , the regret of TS is lower-bounded by a non-zero term since the coefficient of  $\log T$  converges to a non-zero constant. Therefore, TS with prior  $k \geq 1$  is suboptimal, at least for sufficiently small  $\sigma_2$ , where the same result was found in the Gaussian models [Honda and Takemura, 2014]. From Theorem 3.2, we can obtain the following corollary, which shows the suboptimality of some uniform priors with different parameterizations.

**Corollary 3.3.** For any one-to-one transformations  $g(\mu)$  and  $h(\sigma)$ , if  $\frac{d}{d\mu} g^{-1}(\mu) \propto 1$  and  $\frac{d}{d\sigma} h^{-1}(\sigma) \propto \sigma^{-k}$  hold with some  $k \geq 1$ , then TS with the uniform priors with  $(g(\mu), h(\sigma))$  parameterization,  $\pi_u^{g(\mu), h(\sigma)}$ , cannot achieve the lower bound in (3.2).

*Proof.* The uniform prior with  $(g(\mu), h(\sigma))$  parameterization indicates that  $\pi_u^{g(\mu), h(\sigma)} \propto 1$ . Let us define  $f(\mu, \sigma) = (g^{-1}(\mu), h^{-1}(\sigma))$ . Then, the corresponding prior with  $(\mu, \sigma)$  parameterization can be obtained by multiplying the absolute value of the Jacobian determinant of  $f$ , which is given as  $|\det \nabla f| \cdot \pi_u^{g(\mu), h(\sigma)} = \sigma^{-k}$ . Since  $k \geq 1$  holds from the assumption, the proof follows from Theorem 3.2 in this section for the uniform models.  $\square$

Although one might think of such reparameterizations as an artificial example, it is worth noting that the uniform prior with location-rate parameterization, where  $g(\mu) = \mu$  and  $h(\sigma) = \sigma^{-1}$ , coincides with the prior  $k = 2$  with the LS parameterization, which is found to be suboptimal in the uniform model.

Theorem 3.4 below shows the asymptotic optimality of TS-T for the uniform models with the prior with any  $k$ , including the reference prior and the Jeffreys prior that are invariant under any one-to-one transformations.

**Theorem 3.4.** *With the same notation as Theorem 3.1, the expected regret of TS-T with prior  $k \in \mathbb{R}$  for the uniform models is bounded as*

$$\mathbb{E}[\text{Reg}(T)] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{\log \left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)} + \mathcal{O}\left(\epsilon^{-\max(2, k+1)}\right).$$

Note that the remainder term for the Jeffreys prior is larger than that of other optimal priors, including the reference prior, which would stem from the deficiency of the Jeffreys prior under the multiparameter models. Although Theorem 3.4 states that any prior  $\sigma^{-k}$  can achieve the optimal bound *asymptotically*, we recommend using the reference prior  $k = 1$  since small  $k$  requires many initial plays from  $n_0 = \max(2, 3 - \lceil k \rceil)$ , while large  $k$  will suffer from a large regret in the finite time, which is hidden in the remainder term.

### 3.5 Gaussian Bandits

The theoretical findings of TS for the uniform bandits with noninformative priors show significant similarities to those for the Gaussian bandits. This raises a natural question about the performance of TS-T in the Gaussian bandits. In this section, we show the asymptotic optimality of TS-T for the Gaussian bandits with the reference prior and the Jeffreys prior, which was found to be suboptimal with TS.

For the Gaussian bandits, the lower bound in (2.1) can be written as follows [Honda and Takemura, 2014]:

$$\begin{aligned} \liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} &\geq \sum_{i=2}^K \frac{\Delta_i}{\inf_{(\mu, \sigma): \mu > \mu_1} \text{KL}(\text{Gaussian}(\mu_i, \sigma_i); \text{Gaussian}(\mu, \sigma))}, \\ &= \sum_{i=2}^K \frac{\Delta_i}{\frac{1}{2} \log \left(1 + \left(\frac{\Delta_i}{\sigma_i}\right)^2\right)}. \end{aligned} \quad (3.6)$$

Note that the Gaussian distributions satisfy the regularity condition. Therefore, there exist unique maximum likelihood estimators (MLEs) of  $(\mu_i, \sigma_i)$ , denoted by  $(\hat{\mu}_{i,n}, \hat{\sigma}_{i,n})$ , which are written as a function of  $T(X_{i,n}) = (\hat{x}_{i,n}, S_{i,n})$  where  $x_{i,s} \sim \text{Gaussian}(\mu_i, \sigma_i)$ ,

$$\hat{x}_{i,n} = \frac{1}{n} \sum_{s=1}^n x_{i,s}, \quad \text{and} \quad S_{i,n} = \sum_{s=1}^n (x_{i,s} - \hat{x}_{i,n})^2.$$

Their sampling distributions are well-known as follows:

$$\hat{x}_{i,n} \sim \text{Gaussian}\left(\mu_i, \frac{\sigma_i^2}{n}\right), \quad \frac{S_{i,n}}{\sigma_i^2} \sim \chi_{n-1}^2, \quad (3.7)$$

where  $\chi_{n-1}^2$  denotes the chi-squared distribution with degree of freedom  $n - 1$ .

### 3.5.1 Thompson sampling for the Gaussian bandits

The marginal posterior distribution of  $\mu$  under the priors  $\sigma^{-k}$  is given as

$$\begin{aligned}\pi^k(\mu|\hat{\mu}_{i,n}, \hat{\sigma}_{i,n}) &= \frac{\Gamma\left(\frac{n}{2}\right) \left(1 + \left(\frac{\mu - \hat{\mu}_{i,n}}{\hat{\sigma}_{i,n}}\right)^2\right)^{-\frac{n+k-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right) \sqrt{2\pi} \hat{\sigma}_{i,n}} \\ &= f_{n+k-2}^t(\mu; \hat{\mu}_{i,n}, \hat{\sigma}_{i,n}),\end{aligned}\tag{3.8}$$

where  $f_{n+k-2}^t(\cdot; \hat{\mu}_{i,n}, \hat{\sigma}_{i,n})$  denotes the density function of the non-standardized Student's t-distribution with the degree of freedom  $n + k - 2$ , the location  $\hat{\mu}_{i,n} = \hat{x}_{i,n}$ , and the scale  $\hat{\sigma}_{i,n} = \sqrt{S_{i,n}/n}$ . Note that  $\hat{\mu}_{i,n}$  and  $\hat{\sigma}_{i,n}$  are unique MLEs of  $\mu$  and  $\sigma$ , respectively and we require  $n_0 = \max(2, 3 - \lceil k \rceil)$  initial plays to avoid improper posteriors. In the Gaussian models, we can easily sample the location parameter directly from its marginal posterior distribution as it can be expressed by a well-known probability distribution. If one considers the sequential sampling in a similar manner to the case of the uniform models, then one has to sample  $\sigma$  from the inverse transformed gamma distribution and then  $\mu$  from the Gaussian distribution. Note that the marginal posterior of  $\sigma^2$  is the inverse gamma distribution. Since it induces unnecessary additional computation, one would sample  $\tilde{\mu}$  directly from its marginal posterior in the Gaussian bandits.

The following lemma shows the suboptimality of TS with the reference prior ( $k = 1$ ) and the Jeffreys prior ( $k = 2$ ), which is a similar result to Theorem 3.2 in this thesis for the Gaussian bandits.

**Lemma 3.5** (Simplified version of Theorem 2 in Honda and Takemura [2014]). *There exist some Gaussian bandit instances  $\nu$  such that TS with prior  $\sigma^{-k}$  cannot achieve the lower regret bound in (3.6) for  $k \geq 1$ .*

Therefore, Corollary 3.3 can be extended to the Gaussian bandits, which makes the optimality of TS with the uniform prior less informative in practice. This highlights the importance of considering the optimality of TS with invariant priors, which may improve the applicability of TS in practical settings.

**Corollary 3.6.** *For any one-to-one transformations  $g(\mu)$  and  $h(\sigma)$ , if  $\frac{d}{d\mu}g^{-1}(\mu) \propto 1$  and  $\frac{d}{d\sigma}h^{-1}(\sigma) \propto \sigma^{-k}$  hold with some  $k \geq 1$ , then TS with the uniform priors under  $(g(\mu), h(\sigma))$  parameterization,  $\pi_u^{g(\mu), h(\sigma)}$ , cannot achieve the lower bound in (3.6).*

*Proof.* The proof of Corollary 3.3 with Lemma 3.5 concludes the proof.  $\square$

### 3.5.2 Thompson sampling with truncation for the Gaussian bandits

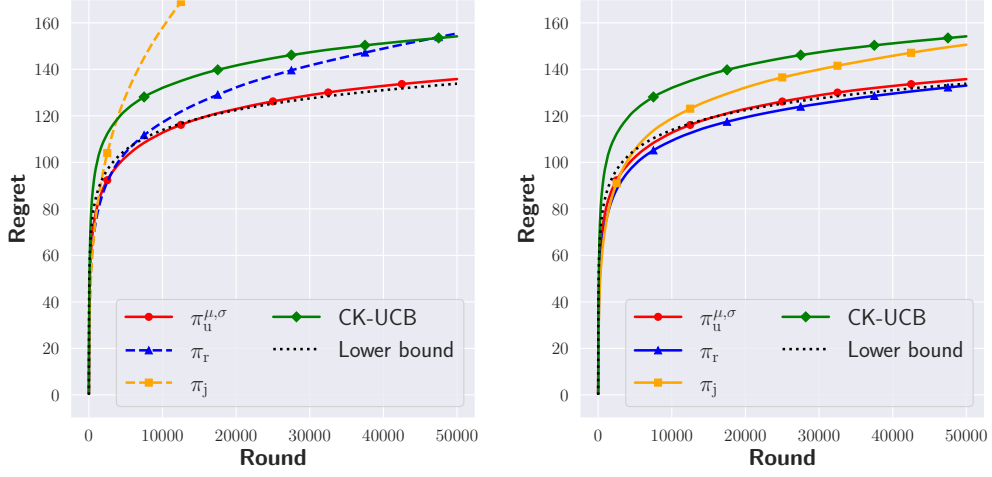
For the realization of the TS-T policy in the Gaussian bandits, we consider a truncated statistic and the corresponding scale estimator as follows:

$$\bar{S}_{i,n} = \max(1, S_{i,n}) \implies \bar{\sigma}_{i,n} = \sqrt{\bar{S}_{i,n}n^{-1}} \geq n^{-\frac{1}{2}}.$$

This implies that TS-T draws a sample of the location parameter from the distribution whose density function is given as

$$\bar{\pi}_{i,n}^k(\mu) = \pi^{G,k}(\mu|\hat{\mu}_{i,n}, \bar{\sigma}_{i,n}) = f_{n+k-2}^t(\mu|\hat{\mu}_{i,n}, \bar{\sigma}_{i,n}),$$

where we just replaced  $\hat{\sigma}_{i,n}$  with  $\bar{\sigma}_{i,n}$  in (3.8). Similarly to Theorem 3.4, TS-T with the reference prior and the Jeffreys prior are also asymptotically optimal in terms of the lower bound in (3.6) for the Gaussian models.



(a) Regret of TS.

(b) Regret of TS-T.

Figure 3.2: Cumulative regret for the 6-armed uniform bandit instance  $\nu_6^U$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively.

**Theorem 3.7.** Assume that the arm 1 is the unique optimal arm with a finite mean. Given arbitrary  $\epsilon \in \left(0, \min_{i \neq 1} \frac{\Delta_i}{2}\right)$ , the expected regret of TS-T with priors  $\sigma^{-k}$  for the Gaussian models is bounded as for  $k \leq 2$

$$\mathbb{E}[\text{Reg}(T)] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{\frac{1}{2} \log \left(1 + \frac{(\Delta_i - 2\epsilon)^2}{\sigma_i^2 + \epsilon}\right)} + \mathcal{O}(\epsilon^{-m}),$$

where  $m = 4 + \lceil k \rceil \mathbb{1}[k \in [1, 2]]$  and  $\lceil \cdot \rceil$  denotes the ceiling function.

Similarly to the uniform models, letting  $\epsilon = o(1)$  in Theorem 3.7 shows the asymptotic optimality of TS-T in the Gaussian models and letting  $\epsilon := \mathcal{O}((\log T)^{-1/7})$  provides an  $\epsilon$ -free bound. Note that the remainder term of TS-T for  $k \in [1, 2]$  is  $\mathcal{O}(\epsilon^{-5+\lceil k \rceil})$ , which is larger than that of TS with the uniform prior,  $\mathcal{O}(\epsilon^{-5})$ , in Honda and Takemura [2014]. The larger remainder term in TS-T can be seen as a cost of introducing a truncation procedure that makes suboptimal priors under TS achieve the optimal regret bound under TS-T. However, we expect that one can find a tighter bound, with  $m = 4 + k \mathbb{1}[k \geq 1]$ , by modifying Lemma 3.21 on the hyperconfluent geometric function of the second kind in Section 3.7.7 or by considering  $\tilde{S}_{i,n} = \max(\frac{1}{n}, S_{i,n})$  such that  $\bar{\sigma}_{i,n} \geq \frac{1}{n}$ .

### 3.6 Simulation Results

This section aims to demonstrate the performance of two policies, TS and TS-T, for the uniform bandits and the Gaussian bandits with different noninformative priors through simulation results. To provide a baseline for comparison, we present the results of asymptotically optimal UCB-based policies, CK-UCB for the uniform bandits [Cowan and Katehakis, 2015] and CHK-UCB for the Gaussian bandits [Cowan et al., 2017]. For the uniform bandits, we consider a 6-armed instance  $\nu_6^U$ , which was previously studied [Cowan and Katehakis, 2015]. Similarly, for the Gaussian bandits, we consider a 6-armed instance  $\nu_6^G$ , which was studied in Cowan et al. [2017]. The values of the parameters used in this section are provided in Table 3.1.

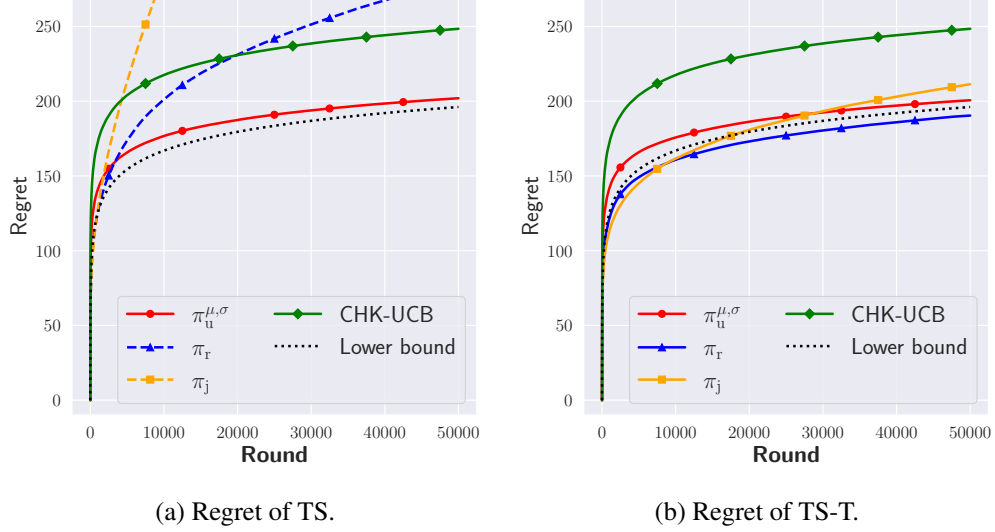


Figure 3.3: Cumulative regret for the 6-armed Gaussian bandit instance  $\nu_6^G$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively.

Table 3.1: Parameters of the 6-armed bandit instances.

(a) Uniform bandit instance $\nu_6^U$ .							(b) Gaussian bandit instance $\nu_6^G$ .						
	Arm 1	Arm 2	Arm 3	Arm 4	Arm 5	Arm 6		Arm 1	Arm 2	Arm 3	Arm 4	Arm 5	Arm 6
$\mu$	5.5	5.0	4.5	4.0	4.75	3.0	$\mu$	10	9	8	7	-1	0
$\sigma$	4.5	5.0	4.5	4	3.75	2.0	$\sigma$	$2\sqrt{2}$	1	1	$\sqrt{0.5}$	1	2

In this section, the solid lines denote the averaged regret over 10,000 independent runs of the policy that was found to be optimal in terms of the regret lower bound in (3.2), whereas the dashed lines denote that of the suboptimal policies. The dotted lines denote the asymptotic regret lower bound. Note that the Jeffreys prior ( $k = 2$ ) coincides with the uniform prior with the location-rate parameterizations ( $l, \sigma^{-1}$ ).

**Uniform bandits** In Figure 3.2a, TS with the uniform prior  $\pi_u^{\mu, \sigma}$  shows the best performance, while TS with the Jeffreys prior  $\pi_j$  and the reference prior  $\pi_r$  suffer from a large regret. Although TS with the reference prior shows a similar performance to CK-UCB, it seems to have a larger order of regret compared to other asymptotically optimal policies. However, as shown in Figure 3.2b, the performance of TS-T with the reference prior is superior, which highlights the effectiveness of the truncation procedure in the TS-based policy.

**Gaussian bandits** Based on the theoretical results of TS and TS-T in Sections 3.4 and 3.5, one can expect that TS and TS-T in the Gaussian bandits will show a similar tendency to that in simulations of the uniform bandits. As we expect, TS with the uniform prior  $\pi_u^{\mu, \sigma}$  shows the best performance, while TS with two invariant priors shows the suboptimal performance in Figure 3.3a. One can see the similar behavior of TS-T in the Gaussian bandits in Figure 3.3b, where the reference prior shows the best performance.



### 3.7 Proofs of Theoretical Results

In this section, we provide detailed proofs of all the theorems and lemmas presented in this chapter.

#### 3.7.1 Derivation of the posteriors for the uniform model

Here, we provide the detailed derivation of the posteriors based on the priors  $\sigma^{-k}$ . Let  $X_n = (x_1, \dots, x_n)$  denote the  $n$  observations of an arm. Then, it holds that

$$\prod_{s=1}^n f_{\mu, \sigma}(x_s) = \frac{1}{\sigma^n} \mathbb{1} \left[ \mu - \frac{\sigma}{2} \leq x^{(1)} \leq x^{(n)} \leq \mu + \frac{\sigma}{2} \right],$$

where  $x^{(1)} = \min_{s \in [n]} x_s$  and  $x^{(n)} = \max_{s \in [n]} x_s$  denotes the smallest and the largest order statistics. Since it holds that

$$\begin{aligned} \mathbb{1} \left[ \mu - \frac{\sigma}{2} \leq x^{(1)} \leq x^{(n)} \leq \mu + \frac{\sigma}{2} \right] \\ = \mathbb{1} \left[ x^{(n)} - \frac{\sigma}{2} \leq \mu \leq x^{(1)} + \frac{\sigma}{2} \right] \mathbb{1}[\sigma \geq x^{(n)} - x^{(1)}], \end{aligned}$$

one can obtain for  $\hat{\sigma}_n = x^{(n)} - x^{(1)}$  that

$$\begin{aligned} \iint \frac{1}{\sigma^k} \prod_{s=1}^n f_{\mu, \sigma}(x_s) d\mu d\sigma &= \int_{\hat{\sigma}_n}^{\infty} \frac{\sigma - \hat{\sigma}_n}{\sigma^{n+k}} d\sigma \\ &= \frac{1}{(n+k-1)(n+k-2)} \frac{1}{(\hat{\sigma}_n)^{n+k-2}}. \end{aligned}$$

Therefore, the joint posterior density can be written as

$$\pi^k(\mu, \sigma \mid X_n) = (n+k-1)(n+k-2) \frac{(\hat{\sigma}_n)^{n+k-2}}{\sigma^{n+k}} \mathbb{1} \left[ \mu - \frac{\sigma}{2} \leq x^{(1)} \leq x^{(n)} \leq \mu + \frac{\sigma}{2} \right].$$

By marginalizing with respect to  $\mu$ , one can obtain the marginal posterior density of  $\sigma$ ,

$$\pi^k(\sigma \mid X_n) = (n+k-1)(n+k-2) \frac{(\hat{\sigma}_n)^{n+k-2}}{\sigma^{n+k}} (\sigma - \hat{\sigma}_n) \mathbb{1}[\sigma \geq x^{(n)} - x^{(1)}].$$

Then, the conditional posterior density of  $\mu$  is written as

$$\begin{aligned} \pi^k(\mu \mid X_n, \sigma) &= \frac{\pi^k(\mu, \sigma \mid X_n)}{\pi^k(\sigma \mid X_n)} \\ &= \frac{1}{\sigma - \hat{\sigma}_n} \mathbb{1} \left[ x^{(n)} - \frac{\sigma}{2} \leq \mu \leq x^{(1)} + \frac{\sigma}{2} \right], \end{aligned}$$

which is the probability density of  $\text{Uniform}_{ab}(x^{(n)} - \sigma/2, x^{(1)} + \sigma/2)$ .

#### 3.7.2 Proof of main results

To begin, we provide a general proof outline that applies to our analysis of TS and TS-T in both the uniform bandits and the Gaussian bandits. Since the overall proofs of Theorems 3.1, 3.4, and 3.7 have a similar structure, we provide the general proof outline here. The proof of Theorem 3.2 is postponed to Section 3.7.9.

At the round  $t$ , we denote the best arm under the posterior sample by  $\tilde{\mu}^*(t) = \max_{i \in [K]} \tilde{\mu}_i(t)$ , which is computed as the maximum of the sampled expected rewards of

all  $K$  arms at round  $t$ . We use the notation  $\mathcal{M}_\epsilon(t)$  to denote an event related to  $\tilde{\mu}^*(t)$  at round  $t$ , which we define for a small positive constant  $\epsilon$  as

$$\mathcal{M}_\epsilon(t) = \{\tilde{\mu}^*(t) \geq \mu_1 - \epsilon\}.$$

Then, the proof starts by decomposing the regret as follows:

$$\begin{aligned} \text{Reg}(T) &= \sum_{t=1}^T \Delta_{i(t)} = \sum_{i=2}^K \Delta_i \mathbb{1}[i(t) = i] \\ &= \sum_{i=2}^K \Delta_i n_0 + \underbrace{\sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1}[i(t) = i, \mathcal{M}_\epsilon^c(t)]}_{\text{bad optimal (BO) term}} \\ &\quad + \underbrace{\sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1}[i(t) = i, \mathcal{M}_\epsilon(t)]}_{\text{good optimal (GO) term}}, \end{aligned} \quad (3.9)$$

where  $n_0$  and the superscript “ $c$ ” denote the number of initial plays and the complementary set, respectively.

(BO) controls the regret induced when the sampled mean parameter of the optimal arm is less than its true value, and (GO) contains the exploration term that becomes the main regret term. Note that (BO) is the main difficulty term of the regret analysis of TS in many bandit models.

The lemmas below conclude the proof of Theorems 3.1 and 3.4, which shows the asymptotic optimality of TS and TS-T for the uniform models with unknown supports.

**Lemma 3.8.** *For the  $K$ -armed uniform bandit models, it holds under both TS and TS-T that*

$$\mathbb{E}[(\text{GO})] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{\log \left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)} + \mathcal{O}(1) + \mathcal{O}(\epsilon^{-1}).$$

**Lemma 3.9.** *For the  $K$ -armed uniform bandit models, it holds that*

$$\mathbb{E}[(\text{BO})] \leq \begin{cases} \mathcal{O}(\epsilon^{-2}) & \text{under TS with prior } \sigma^{-k}, k < 1, \\ \mathcal{O}(\epsilon^{-\max(2, k+1)}) & \text{under TS-T with prior } \sigma^{-k}, k \in \mathbb{R}. \end{cases}$$

In the proof of Lemma 3.9, our analysis cannot derive the finite upper-bound for TS with  $k \geq 1$ , including the reference prior and the Jeffreys prior, where the same problem was observed in the Gaussian models [Honda and Takemura, 2014]. Although the infinite upper-bound term does not necessarily mean the suboptimality of the policy, Theorem 3.2 shows that it actually contributes to increasing the regret in expectation. This is because TS could induce a polynomial regret with a small but non-negligible probability, which leads to a larger expected regret. A truncation procedure is introduced to make such a probability ignorable so that (BO) can be upper-bounded by a finite term.

The lemmas below conclude the proof of Theorem 3.7, which shows the asymptotic optimality of TS-T for the Gaussian bandits.

**Lemma 3.10.** *For the  $K$ -armed Gaussian bandit models, it holds under TS-T that*

$$\mathbb{E}[(\text{GO})] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{\frac{1}{2} \log \left(1 + \frac{(\Delta_i - 2\epsilon)^2}{\sigma_i^2 + \epsilon}\right)} + \mathcal{O}(1) + \mathcal{O}(\epsilon^{-2}).$$

**Lemma 3.11.** *For the  $K$ -armed Gaussian bandit models, it holds under TS-T with prior  $k \leq 2$  that*

$$\mathbb{E}[(\text{BO})] \leq \mathcal{O}(\epsilon^{-m}),$$

where  $m = 4 + \lceil k \rceil \mathbb{1}[k \in [1, 2]]$  and  $\lceil \cdot \rceil$  denotes the ceiling function.

Although some parts of the proof of Lemmas 3.10 and 3.11 can be obtained using the results by Honda and Takemura [2014], the main difficulty comes from the term introduced by the truncated estimators  $\bar{\sigma}$ . To be precise, it involves integrating a product of the beta function and the incomplete gamma function. This integration introduces additional functions, such as the modified Bessel function of the second kind and the confluent hypergeometric function of the second kind, which makes the analysis technically more complicated.

### 3.7.3 Proof of Lemma 3.8

Before beginning the proof, we first introduce the result that demonstrates the joint distribution of the order statistics of the uniform distribution. Here, an additional notation in superscript,  $\text{SD}_U$ , is used to clarify that it is a density function of the sampling distribution in the uniform models.

**Lemma 3.12** (Lemma 6 in Cowan and Katehakis [2015]). *Let  $x_1, x_2, \dots, x_n$  be i.i.d. random variables following  $\text{Uniform}_{ab}(a, b)$ , with finite  $a < b$ . For  $n \geq 2$ , let  $x^{(n)} = \max_{s \in [n]} x_s$  and  $x^{(1)} = \min_{s \in [n]} x_s$ . Then, the joint density of  $(x^{(1)}, x^{(n)})$  is given by*

$$f_n^{\text{SD}_U}(y, z) = \begin{cases} n(n-1)(b-a)^{-n}(z-y)^{n-2} & \text{if } y \leq z, \\ 0 & \text{otherwise.} \end{cases}$$

*Proof.* Recall that we use MLEs,  $\hat{\mu}_i(n) = \frac{x_i^{(1)} + x_i^{(n)}}{2}$  and  $\hat{\sigma}_i(n) = x_i^{(n)} - x_i^{(1)}$ , which are functions of sufficient statistics  $T(X_{i,n}) = (x_i^{(1)}, x_i^{(n)})$ . Define events on the order statistic  $x_i^{(1)}$  and  $x_i^{(n)}$ , and an event on the truncated statistic  $\bar{x}_i^{(n)}$  of the arm  $i$  at round  $t$  for any positive  $\epsilon < \frac{\Delta_i}{2}$ ,

$$\begin{aligned} \mathcal{A}_{i,n}(\epsilon) &= \left\{ \mu_i - \frac{\sigma_i}{2} \leq x_i^{(1)} \leq \mu_i - \frac{\sigma_i}{2} + \epsilon \right\} \\ \mathcal{B}_{i,n}(\epsilon) &= \left\{ \mu_i + \frac{\sigma_i}{2} - \epsilon \leq x_i^{(n)} \leq \mu_i + \frac{\sigma_i}{2} \right\} \\ \mathcal{E}_{i,n}(\epsilon) &= \mathcal{A}_{i,n}(\epsilon) \cap \mathcal{B}_{i,n}(\epsilon) \\ \bar{\mathcal{B}}_{i,n}(\epsilon) &= \left\{ \mu_i + \frac{\sigma_i}{2} - \epsilon \leq \bar{x}_i^{(n)} \leq \mu_i + \frac{\sigma_i}{2} \right\} \\ \bar{\mathcal{E}}_{i,n}(\epsilon) &= \mathcal{A}_{i,n}(\epsilon) \cap \bar{\mathcal{B}}_{i,n}(\epsilon). \end{aligned}$$

Then, (GO) is decomposed under TS by

$$\begin{aligned} (\text{GO}) &= \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] \\ &\quad + \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}^c(\epsilon), \mathcal{M}_\epsilon(t)] \\ &\leq \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] + \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}^c(\epsilon)]. \end{aligned}$$

The last equality holds since an event  $\{i(t) = i, N_i(t) = n\}$  occurs only once from the definition  $N_i(t)$ . Similarly, (GO) can be decomposed under TS-T by

$$\begin{aligned} (\text{GO}) &\leq \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1} [i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] \\ &\quad + \Delta_i \mathbb{1} [i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}^c(\epsilon), ]. \end{aligned}$$

Then, two lemmas below conclude the proof of Lemma 3.8, whose proofs are postponed to Section 3.7.5.  $\square$

**Lemma 3.13.** *For all  $i \in [K]$  and  $n \in \mathbb{N}_{\geq 2}$ , it holds that*

$$\mathbb{P}[\bar{\mathcal{E}}_{i,n}^c(\epsilon)] \leq \mathbb{P}[\mathcal{E}_{i,n}^c(\epsilon)] \leq 2 \exp\left(-\frac{\epsilon}{\sigma_i} n\right).$$

**Lemma 3.14.** *Under TS, it holds that for any  $i \in [K]$  and given  $\epsilon \in (0, \frac{\Delta_i}{2})$*

$$\sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = i, \mathcal{E}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)]] \leq \max\left(\frac{1}{2}, \frac{5}{2} - k\right) + \frac{\log T}{\log\left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)}.$$

and under TS-T,

$$\sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)]] \leq \frac{3}{2} + \frac{1}{\sigma_i} + \frac{\log T}{\log\left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)}.$$

### 3.7.4 Proof of Lemma 3.9

*Proof.* Let us consider the following decomposition:

$$\begin{aligned} &\sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{M}_\epsilon^c(t)] \\ &= \sum_{n=n_0}^T \sum_{t=Kn_0+1}^T \mathbb{1}\left[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n\right] \\ &= \sum_{n=n_0}^T \sum_{m=1}^T \mathbb{1}\left[m \leq \sum_{t=Kn_0+1}^T \mathbb{1}\left[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n\right]\right]. \end{aligned}$$

Notice that

$$m \leq \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n]$$

implies that  $\tilde{\mu}_1(t) \leq \max_{i \in [K]} \tilde{\mu}_i(t) \leq \mu_1 - \epsilon$  occurred  $m$  times in a row on  $\{t : \mathcal{M}_\epsilon^c(t), N_1(t) = n\}$ . Therefore, we obtain that

$$(\text{BO}) \leq \sum_{i=2}^K \sum_{n=n_0}^T \sum_{m=1}^T \Delta_i \mathbb{1}\left[m \leq \sum_{t=Kn_0+1}^T \mathbb{1}\left[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n\right]\right].$$

Firstly, we provide the upper bound of  $\mathbb{E}[(\text{BO})]$  under TS.

### Under TS

Let us define  $p_n(y|\theta_{1,n}) = \mathbb{P} \left[ \tilde{\mu}_1 \geq \mu_1 - y \mid x_1^{(1)}, x_1^{(n)} \right]$ , where  $p_n(\epsilon|\theta_{1,n})$  denote the probability that  $\mathcal{M}_\epsilon(t)$  occurs given sufficient statistics  $\theta_{1,n} = T(X_{i,n})$  where  $N_1(t) = n$ . Therefore, we have

$$\begin{aligned} \mathbb{E}[(\text{BO})] &\leq \sum_{i=2}^K \sum_{n=n_0}^T \sum_{m=1}^T \Delta_i \mathbb{P} \left[ m \leq \sum_{t=Kn_0+1}^T \mathbb{1} \left[ i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n \right] \right] \\ &\leq \sum_{i=2}^K \sum_{n=n_0}^T \mathbb{E}_{\theta_{1,n}} \left[ \sum_{m=1}^T (1 - p_n(\epsilon|\theta_{1,n}))^m \right] \\ &\leq \sum_{n=n_0}^T \mathbb{E}_{\theta_{1,n}} \left[ \frac{1 - p_n(\epsilon|\theta_{1,n})}{p_n(\epsilon|\theta_{1,n})} \right], \end{aligned} \quad (3.10)$$

where we utilized the total law of expectation in (3.10). From now on, we fix  $n$  so that  $(\tilde{\mu}_1, \tilde{\sigma}_1)$  are sampled from the same posterior parameterized by fixed  $(\hat{\mu}_{1,n}, \hat{\sigma}_{1,n})$  and drop the subscript  $\theta_{1,n}$  of  $\mathbb{E}_{\theta_{1,n}}$  for simplicity. Therefore,  $\tilde{\sigma}_1 \geq \hat{\sigma}_{1,n} = x_1^{(n)} - x_1^{(1)}$  holds from its marginal posterior in (3.3), which implies the existence of a positive random variable  $D$  satisfying  $\tilde{\sigma}_1 = x_1^{(n)} - x_1^{(1)} + D$ . From the sequential sampling in Algorithm 2, it holds that  $\tilde{\mu}_1 \sim \text{Uniform}_{\mu\sigma} \left( \frac{x_1^{(1)} + x_1^{(n)}}{2}, D \right)$ . Therefore, if  $\frac{x_1^{(1)} + x_1^{(n)}}{2} \geq$

$\mu_1 - \epsilon$ ,  $p_n(\epsilon|\theta_{1,n}) \geq \frac{1}{2}$  holds regardless the value of  $D$ . Since  $\frac{x_1^{(1)} + x_1^{(n)}}{2} \geq \mu_1 - \frac{\epsilon}{2}$  holds on  $\mathcal{E}_{1,n}(\epsilon)$ , we obtain

$$\begin{aligned} \mathbb{E} \left[ \frac{1 - p_n(\epsilon|\theta_{1,n})}{p_n(\epsilon|\theta_{1,n})} \right] &\leq 2\mathbb{E} \left[ \mathbb{1} \left[ x_1^{(1)} + x_1^{(n)} \geq 2(\mu_1 - \epsilon) \right] (1 - p_n(\epsilon|\theta_{1,n})) \right] \\ &\quad + \mathbb{E} \left[ \frac{\mathbb{1} \left[ x_1^{(1)} + x_1^{(n)} \leq 2(\mu_1 - \epsilon) \right]}{p_n(\epsilon|\theta_{1,n})} \right] \\ &\leq 2\mathbb{P} \left( \mathcal{E}_{1,n}^c \left( \frac{\epsilon}{2} \right) \right) + 2\mathbb{E} \left[ \mathbb{1} \left[ \mathcal{E}_{1,n} \left( \frac{\epsilon}{2} \right) \right] (1 - p_n(\epsilon|\theta_{1,n})) \right] \\ &\quad + \mathbb{E} \left[ \frac{\mathbb{1} \left[ x_1^{(1)} + x_1^{(n)} \leq 2(\mu_1 - \epsilon) \right]}{p_n(\epsilon|\theta_{1,n})} \right]. \end{aligned} \quad (3.11)$$

From Lemma 3.13, the first term of (3.11) can be bounded as

$$2\mathbb{P} \left( \mathcal{E}_{1,n}^c \left( \frac{\epsilon}{2} \right) \right) \leq 4e^{-\frac{\epsilon}{2\sigma_1}n}. \quad (3.12)$$

Since  $\tilde{\mu}_1|\tilde{\sigma}_1 \sim \text{Uniform}_{ab} \left( x_1^{(n)} - \frac{\tilde{\sigma}_1}{2}, x_1^{(1)} + \frac{\tilde{\sigma}_1}{2} \right)$ , we have

$$x_1^{(n)} - \frac{\tilde{\sigma}_1}{2} \geq \mu_1 - \epsilon \Leftrightarrow \tilde{\sigma}_1 \leq 2(x_1^{(n)} - (\mu_1 - \epsilon)) \implies 1 - p_n(\epsilon|\theta_{1,n}) = 0.$$

For a constant  $A = x_1^{(n)} - (\mu_1 - \epsilon)$ , we can bound the second term of (3.11) as

$$\begin{aligned}
& \mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] \left( 1 - p_n(\epsilon|\theta_{1,n}) \right) \\
&= \mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] \int_{2A}^{\infty} \pi^k(s|\theta_{1,n}) \int_{x_1^{(n)} - \frac{s}{2}}^{\mu_1 - \epsilon} \pi^k(m|\theta_{1,n}, \tilde{\sigma}_1 = s) dm ds \\
&= \mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] \int_{2A}^{\infty} \frac{(n+k-1)(n+k-2) \left( x_1^{(n)} - x_1^{(1)} \right)^{n+k-2}}{s^{n+k}} \\
&\quad \cdot \left( s - (x_1^{(n)} - x_1^{(1)}) \right) \int_{x_1^{(n)} - \frac{s}{2}}^{\mu_1 - \epsilon} f_{\hat{\mu}_{1,n}, s - \hat{\sigma}_{1,n}}^{\text{U}\mu\sigma}(m) dm ds \\
&= \mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] \int_{2A}^{\infty} \frac{(n+k-1)(n+k-2) \left( x_1^{(n)} - x_1^{(1)} \right)^{n+k-2}}{s^{n+k}} \\
&\quad \cdot \left( s - (x_1^{(n)} - x_1^{(1)}) \right) \frac{1}{s - x_1^{(n)} - x_1^{(1)}} \left( \frac{s - 2A}{2} \right) ds \\
&= \mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] \int_{2A}^{\infty} \frac{(n+k-1)(n+k-2) \left( x_1^{(n)} - x_1^{(1)} \right)^{n+k-2}}{s^{n+k}} \left( \frac{s - 2A}{2} \right) ds \\
&= \mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] \frac{1}{2} \left( \frac{x_1^{(n)} - x_1^{(1)}}{2(x_1^{(n)} - \mu_1 + \epsilon)} \right)^{n+k-2}.
\end{aligned}$$

Since  $x_1^{(n)} \geq \mu_1 + \frac{\sigma_1}{2} - \frac{\epsilon}{2}$  and  $x_1^{(n)} - x_1^{(1)} \leq \sigma_1$  hold for any  $n$  on  $\mathcal{E}_{1,n}(\epsilon/2)$ , we have

$$\begin{aligned}
\mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] (1 - p_n(\epsilon|\theta_{1,n})) &\leq \mathbb{1} \left[ \mathcal{E}_{1,n}(\epsilon/2) \right] \frac{1}{2} \left( \frac{x_1^{(n)} - x_1^{(1)}}{2(x_1^{(n)} - \mu_1 + \epsilon)} \right)^{n+k-2} \\
&\leq \frac{1}{2} \left( \frac{\sigma_1}{\sigma_1 + \epsilon} \right)^{n+k-2} \leq \frac{1}{2} e^{-\frac{\epsilon}{\sigma_1 + \epsilon}(n+k-2)}.
\end{aligned}$$

Therefore, the second term of (3.11) is bounded as

$$2\mathbb{E} \left[ \mathbb{1} \left[ \mathcal{E}_{1,n} \left( \frac{\epsilon}{2} \right) \right] (1 - p_n(\epsilon|\theta_{1,n})) \right] \leq e^{-\frac{\epsilon}{\sigma_1 + \epsilon}(n+k-2)}. \quad (3.13)$$

Finally, we evaluate the last term of (3.11). Again, from the conditional posterior of  $\mu$ , we have  $\tilde{\mu}_1|\tilde{\sigma}_1 \sim \text{Uniform}_{ab} \left( x_1^{(n)} - \frac{\tilde{\sigma}_1}{2}, x_1^{(1)} + \frac{\tilde{\sigma}_1}{2} \right)$ , which gives that

$$\begin{aligned}
& \mathbb{1} \left[ x_1^{(1)} + x_1^{(n)} \leq 2(\mu - \epsilon) \right] \mathbb{P}[\tilde{\mu}_1 \geq \mu_1 - \epsilon | \theta_{1,n}, \sigma = \tilde{\sigma}_1] \\
&= \mathbb{1} \left[ x_1^{(1)} + x_1^{(n)} \leq 2(\mu_1 - \epsilon) \right] \cdot \begin{cases} 0 & \text{if } x_1^{(1)} + \frac{\tilde{\sigma}_1}{2} \leq \mu_1 - \epsilon, \\ \frac{x_1^{(1)} + \tilde{\sigma}_1/2 - (\mu_1 - \epsilon)}{\tilde{\sigma}_1 - (x_1^{(n)} - x_1^{(1)})} & \text{otherwise.} \end{cases}
\end{aligned}$$

For simplicity in notation, we denote the event  $\{x_1^{(1)} + x_1^{(n)} \leq 2(\mu_1 - \epsilon)\}$  by  $\mathcal{T}$ . For a constant  $A' = \mu_1 - \epsilon - x_1^{(1)}$ , it holds that

$$\begin{aligned}
& \mathbb{1} \left[ x_1^{(1)} + x_1^{(n)} \leq 2(\mu_1 - \epsilon) \right] p_n(\epsilon | \theta_{1,n}) \\
&= \mathbb{1} [\mathcal{T}] \int_{2(\mu_1 - \epsilon - x_1^{(1)})}^{\infty} \pi^k(s | \theta_{1,n}) \frac{x_1^{(1)} + s/2 - (\mu_1 - \epsilon)}{s - (x_1^{(n)} - x_1^{(1)})} ds \\
&= \mathbb{1} [\mathcal{T}] \int_{2A'}^{\infty} \frac{(n+k-1)(n+k-2) (x_1^{(n)} - x_1^{(1)})^{n+k-2}}{s^{n+k}} \\
&\quad \cdot \left( s - (x_1^{(n)} - x_1^{(1)}) \right) \frac{x_1^{(1)} + \frac{s}{2} - (\mu_1 - \epsilon)}{s - (x_1^{(n)} - x_1^{(1)})} ds \\
&= \mathbb{1} [\mathcal{T}] \int_{2A'}^{\infty} \frac{x_1^{(1)} + \frac{s}{2} - (\mu_1 - \epsilon)}{s^{n+k}} (n+k-1)(n+k-2) (x_1^{(n)} - x_1^{(1)})^{n+k-2} ds \\
&= \mathbb{1} [\mathcal{T}] \frac{(n+k-1)(n+k-2) (x_1^{(n)} - x_1^{(1)})^{n-1}}{2} \int_{2A'}^{\infty} \frac{s - 2A'}{s^{n+k}} ds \\
&= \mathbb{1} [\mathcal{T}] \frac{1}{2} \left( \frac{x_1^{(n)} - x_1^{(1)}}{2A'} \right)^{n+k-2} = \mathbb{1} [\mathcal{T}] \frac{1}{2} \left( \frac{x_1^{(n)} - x_1^{(1)}}{2(\mu_1 - \epsilon - x_1^{(1)})} \right)^{n+k-2}.
\end{aligned}$$

Taking expectations gives that

$$\begin{aligned}
\mathbb{E} \left[ \frac{\mathbb{1} [\mathcal{T}]}{p_n(\epsilon | \theta_{1,n})} \right] &= 2 \underbrace{\mathbb{E} \left[ \mathbb{1} [\mathcal{T}] \left( \frac{2(\mu_1 - \epsilon - x_1^{(1)})}{x_1^{(n)} - x_1^{(1)}} \right)^{n+k-2} \right]}_{(\star_U)} \\
&= 2 \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_y^{\min(2(\mu_1 - \epsilon - y), \mu_1 + \frac{\sigma_1}{2})} f_n^{\text{SD}_U}(y, z) dz dy.
\end{aligned} \tag{3.14}$$

By injecting the sampling distributions of the order statistics in Lemma 3.12, we obtain that

$$\begin{aligned}
(\star_U) &= \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_y^{\min(2(\mu_1 - \epsilon - y), \mu_1 + \frac{\sigma_1}{2})} \frac{n(n-1)}{\sigma_1^n} (z-y)^{n-2} \\
&\quad \cdot \left( \frac{2(\mu_1 - \epsilon - y)}{z-y} \right)^{n+k-2} dz dy.
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
(\star_U) &\leq \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_y^{2(\mu_1 - \epsilon - y)} \frac{n(n-1)}{\sigma_1^n} (z-y)^{n-2} \left( \frac{2(\mu_1 - \epsilon - y)}{z-y} \right)^{n+k-2} dz dy \\
&= \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_y^{2(\mu_1 - \epsilon - y)} \frac{n(n-1)}{\sigma_1^n} \frac{(2(\mu_1 - \epsilon - y))^{n+k-2}}{(z-y)^k} dz dy \\
&= \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \frac{n(n-1)}{\sigma_1^n} (2(\mu_1 - \epsilon - y))^{n+k-2} (2(\mu_1 - \epsilon - y) - y)^{1-k} dy \quad \text{if } k < 1.
\end{aligned} \tag{3.15}$$

Note that under TS with  $k \geq 1$ , the integral in (3.15) with respect to  $z$  becomes infinite due to  $z - y = 0$ , which implies that our analysis does not result in a finite upper bound for  $k \geq 1$ . One can avoid such infinite terms by modifying the domain of the integral with respect to  $z$  from  $[y, 2(mu_1 - \epsilon - y)]$  to  $[y + \alpha, 2(\mu_1 - \epsilon - y)]$  for some  $\alpha > 0$ . Since  $f_n(y, z)$  is the joint density of  $(x^{(1)}, x^{(n)})$ , the domain restriction on  $z$  can be interpreted as an additional restriction  $x^{(n)} \geq x^{(1)} + \alpha$ , which motivates us to design the TS-T policy.

By defining  $w = 2(\mu_1 - \epsilon - y)$ , we can obtain for  $k < 1$  that

$$\begin{aligned}
(\star_U) &\leq \frac{n(n-1)}{2(1-k)\sigma_1^n} \int_0^{\sigma_1-2\epsilon} \left(w - \left(\mu_1 - \epsilon - \frac{w}{2}\right)\right)^{1-k} w^{n+k-2} dw \\
&\leq \frac{n(n-1)}{2(1-k)\sigma_1^n} \int_0^{\sigma_1-2\epsilon} \left(\frac{3w}{2}\right)^{1-k} w^{n+k-2} dw \\
&= \frac{3(n-1)}{4(1-k)\sigma_1^n} (\sigma_1 - 2\epsilon)^n = \frac{3(n-1)}{4(1-k)} \left(1 - \frac{2\epsilon}{\sigma_1}\right)^n \\
&\leq \frac{3(n-1)}{4(1-k)} e^{-\frac{2\epsilon}{\sigma_1}n}.
\end{aligned} \tag{3.16}$$

Therefore, by combining (3.12), (3.13), and (3.16) with (3.11) and (3.10), we have for  $\epsilon \in \left(0, \min_{i \neq 1} \frac{\Delta_i}{2}\right)$  and  $k < 1$  that

$$\begin{aligned}
\sum_{n=n_0}^T \mathbb{E} \left[ \frac{1 - p_n(\epsilon|\theta_{1,n})}{p_n(\epsilon|\theta_{1,n})} \right] &\leq \sum_{n=n_0}^T 4e^{-\frac{\epsilon}{2\sigma}n} + e^{-\frac{\epsilon}{\sigma+\epsilon}(n-1)} + \frac{3(n-1)}{4(1-k)} e^{-\frac{2\epsilon}{\sigma}n} \\
&\leq \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-2}) \\
&= \mathcal{O}(\epsilon^{-2}),
\end{aligned}$$

which concludes the proof of Lemma 3.9 for the case of TS.

### Under TS-T

The overall proofs are the same as that of TS, except we replace  $p_n(\cdot)$  with  $\bar{p}_n(y|\theta_{1,n}) = p_n(y|\bar{\theta}_{1,n}) = \mathbb{P}[\tilde{\mu}_1 \geq \mu_1 - y|x_1^{(1)}, \bar{x}_1^{(n)}]$  and replace  $x^{(n)}$  with  $\bar{x}^{(1)}$ . Similarly to (3.10) in TS, we have for TS-T that

$$\begin{aligned}
\mathbb{E} \left[ \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{M}_\epsilon^c(t)] \right] &\leq \mathbb{E} \left[ \sum_{n=n_0}^T \sum_{m=1}^T (1 - \bar{p}_n(\epsilon|\theta_{1,n}))^m \right] \\
&\leq \sum_{n=n_0}^T \mathbb{E} \left[ \frac{1 - \bar{p}_n(\epsilon|\theta_{1,n})}{\bar{p}_n(\epsilon|\theta_{1,n})} \right].
\end{aligned} \tag{3.17}$$

By following the same steps as (3.11), we obtain that

$$\begin{aligned}
\mathbb{E} \left[ \frac{1 - \bar{p}_n(\epsilon|\theta_{1,n})}{\bar{p}_n(\epsilon|\theta_{1,n})} \right] &\leq 2\mathbb{E} \left[ \mathbb{1} \left[ x_1^{(1)} + \bar{x}_1^{(n)} \geq 2(\mu_1 - \epsilon) \right] (1 - \bar{p}_n(\epsilon|\theta_{1,n})) \right] \\
&\quad + \mathbb{E} \left[ \frac{\mathbb{1} \left[ x_1^{(1)} + \bar{x}_1^{(n)} \leq 2(\mu_1 - \epsilon) \right]}{\bar{p}_n(\epsilon|\theta_{1,n})} \right] \\
&\leq 2\mathbb{P} \left( \bar{\mathcal{E}}_{1,n}^c \left( \frac{\epsilon}{2} \right) \right) + 2\mathbb{E} \left[ \mathbb{1} \left[ \bar{\mathcal{E}}_{1,n} \left( \frac{\epsilon}{2} \right) \right] (1 - \bar{p}_n(\epsilon|\theta_{1,n})) \right] \\
&\quad + \mathbb{E} \left[ \frac{\mathbb{1} \left[ x_1^{(1)} + \bar{x}_1^{(n)} \leq 2(\mu_1 - \epsilon) \right]}{\bar{p}_n(\epsilon|\theta_{1,n})} \right].
\end{aligned} \tag{3.18}$$



From Lemma 3.13, the first term of (3.18) can be bounded as

$$2\mathbb{P}\left(\bar{\mathcal{E}}_{1,n}^c\left(\frac{\epsilon}{2}\right)\right) \leq 4e^{-\frac{\epsilon}{2\sigma_1}n}. \quad (3.19)$$

Since  $\bar{x}_1^{(n)} \geq \mu_1 + \frac{\sigma_1}{2} - \frac{\epsilon}{2}$  and  $\bar{x}_1^{(n)} - x_1^{(1)} \leq \sigma_1$  hold for any  $n$  on  $\bar{\mathcal{E}}_{1,n}(\epsilon/2)$ , we have

$$\begin{aligned} \mathbb{1}\left[\bar{\mathcal{E}}_{1,n}(\epsilon/2)\right] (1 - p_n(\epsilon|\theta_{1,n})) &\leq \mathbb{1}\left[\bar{\mathcal{E}}_{1,n}(\epsilon/2)\right] \frac{1}{2} \left(\frac{\bar{x}_1^{(n)} - x_1^{(1)}}{2(\bar{x}_1^{(n)} - \mu_1 + \epsilon)}\right)^{n+k-2} \\ &\leq \frac{1}{2} \left(\frac{\sigma_1}{\sigma_1 + \epsilon}\right)^{n+k-2} \leq \frac{1}{2} e^{-\frac{\epsilon}{\sigma_1 + \epsilon}(n+k-2)}. \end{aligned}$$

Therefore, the second term of (3.18) is bounded as

$$2\mathbb{E}\left[\mathbb{1}\left[\bar{\mathcal{E}}_{1,n}\left(\frac{\epsilon}{2}\right)\right] (1 - \bar{p}_n(\epsilon|\theta_{1,n}))\right] \leq e^{-\frac{\epsilon}{\sigma_1 + \epsilon}(n+k-2)}. \quad (3.20)$$

Finally, we evaluate the last term of (3.18). By following the same steps to the last term of (3.11), one can obtain for  $\bar{\mathcal{T}} = \left\{x_1^{(1)} + \bar{x}_1^{(n)} \leq 2(\mu - \epsilon)\right\}$

$$\begin{aligned} \mathbb{1}\left[x_1^{(1)} + \bar{x}_1^{(n)} \leq 2(\mu - \epsilon)\right] \mathbb{P}[\tilde{\mu}_1 \geq \mu_1 - \epsilon|\theta_{1,n}, \sigma = \tilde{\sigma}_1] \\ \leq \mathbb{1}[\bar{\mathcal{T}}] \frac{1}{2} \left(\frac{\bar{x}_1^{(n)} - x_1^{(1)}}{2(\mu_1 - \epsilon - x_1^{(1)})}\right)^{n+k-2}. \end{aligned}$$

Since  $\mathbb{1}[\bar{x}_1^{(n)} \neq x_1^{(n)}] = \mathbb{1}\left[\bar{x}_1^{(n)} = x_1^{(1)} + \frac{1}{n}\right]$  from the definition of  $\bar{x}_1^{(n)}$ , taking expectation gives us

$$\begin{aligned} \mathbb{E}\left[\frac{\mathbb{1}[\bar{\mathcal{T}}]}{p_n(\epsilon|\theta_{1,n})}\right] &= 2\mathbb{E}\left[\mathbb{1}\left[x_1^{(1)} + \bar{x}_1^{(n)} \leq 2(\mu_1 - \epsilon)\right] \left(\frac{2(\mu_1 - \epsilon - x_1^{(1)})}{\bar{x}_1^{(n)} - x_1^{(1)}}\right)^{n+k-2}\right] \\ &= 2\mathbb{E}\left[\underbrace{\mathbb{1}\left[\bar{\mathcal{T}}, \bar{x}_1^{(n)} = x_1^{(n)}\right] \left(\frac{2(\mu_1 - \epsilon - x_1^{(1)})}{\bar{x}_1^{(n)} - x_1^{(1)}}\right)^{n+k-2}}_{(\dagger_U)}\right] \\ &\quad + 2\mathbb{E}\left[\underbrace{\mathbb{1}\left[\bar{\mathcal{T}}, \bar{x}_1^{(n)} = x_1^{(1)} + 1/n\right] \left(\frac{2(\mu_1 - \epsilon - x_1^{(1)})}{\bar{x}_1^{(n)} - x_1^{(1)}}\right)^{n+k-2}}_{(\diamond_U)}\right]. \end{aligned}$$

Note that  $(\diamond_U)$  term is introduced due to the truncation procedure in TS-T.

**(1) Upper bound of  $(\dagger_U)$**  Under the condition  $\{x_1^{(1)} + \bar{x}_1^{(n)} \leq 2(\mu_1 - \epsilon), \bar{x}_1^{(n)} = x_1^{(n)}\}$ , we have

$$\begin{aligned} x_1^{(1)} &\in \left[\mu_1 - \frac{\sigma_1}{2}, \mu_1 - \epsilon\right) \\ x_1^{(n)} &\in \left[x_1^{(1)} + \frac{1}{n}, \min(2(\mu_1 - \epsilon - x_1^{(1)}), \mu_1 + \frac{\sigma_1}{2})\right). \end{aligned}$$

By applying Lemma 3.12, we obtain

$$\begin{aligned}
(\dagger_U) &= \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_{y + \frac{1}{n}}^{\min(2(\mu_1 - \epsilon - y), \mu_1 + \frac{\sigma_1}{2})} \frac{n(n-1)}{\sigma_1^n} (z-y)^{n-2} \\
&\quad \cdot \left( \frac{2(\mu_1 - \epsilon - y)}{z-y} \right)^{n+k-2} dz dy \\
&\leq \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_{y + \frac{1}{n}}^{2(\mu_1 - \epsilon - y)} \frac{n(n-1)}{\sigma_1^n} (z-y)^{n-2} \left( \frac{2(\mu_1 - \epsilon - y)}{z-y} \right)^{n+k-2} dz dy \\
&= \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_{y + \frac{1}{n}}^{2(\mu_1 - \epsilon - y)} \frac{n(n-1)}{\sigma_1^n} \frac{(2(\mu_1 - \epsilon - y))^{n+k-2}}{(z-y)^k} dz dy. \tag{3.21}
\end{aligned}$$

Note that the domain of the integral respect to  $z$  in (3.21) is  $[y + n^{-1}, 2(\mu_1 - \epsilon - y)]$  differently from the integral introduced in TS,  $[y, 2(\mu_1 - \epsilon - y)]$  in (3.15). The upper bounds on  $(\star_U)$  in (3.14) directly gives the upper bound of  $(\dagger_U)$  for  $k < 1$ . By defining  $w = 2(\mu_1 - \epsilon - y)$ , we can derive the upper bound of  $(\dagger_U)$  for  $k \geq 1$ .

**(1-i) For the reference prior ( $k = 1$ ):**

$$\begin{aligned}
(\dagger_U) &\leq \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_{y + \frac{1}{n}}^{2(\mu_1 - \epsilon - y)} \frac{n(n-1)}{\sigma_1^n} \frac{(2(\mu_1 - \epsilon - y))^{n-1}}{(z-y)} dz dy \\
&= \frac{1}{2} \int_0^{\sigma_1 - 2\epsilon} \frac{n(n-1)}{\sigma_1^n} w^{n-1} \log(nw) dw \\
&= \frac{1}{2} \left( \frac{\sigma_1 - 2\epsilon}{\sigma_1} \right)^n n(n-1) \frac{n \log(n(\sigma_1 - 2\epsilon)) - 1}{n^2} \\
&\leq \frac{n \log(n\sigma_1)}{2} e^{-\frac{2\epsilon}{\sigma_1} n}.
\end{aligned}$$

**(1-ii) For priors with  $k > 1$ :** One can see that the integral in (3.21) is an increasing function with respect to  $k$  since  $2(\mu_1 - \epsilon - y) > (z-y)$  holds for all  $z \in (y + 1/n, 2(\mu_1 - \epsilon - y))$ . For  $k > 1$ , it holds that

$$\begin{aligned}
(\dagger_U) &\leq \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_{y + \frac{1}{n}}^{2(\mu_1 - \epsilon - y)} \frac{n(n-1)}{\sigma_1^n} \frac{(2(\mu_1 - \epsilon - y))^{n+k-2}}{(z-y)^k} dz dy \\
&\leq \frac{n(n-1)}{\sigma_1^n} \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} (2(\mu_1 - \epsilon - y))^{n+k-2} \frac{n^{k-1}}{k-1} dy \\
&= \frac{n^k(n-1)}{2\sigma_1^n(k-1)} \int_0^{\sigma_1 - 2\epsilon} w^{n+k-2} dw = \frac{n^k(n-1)}{2(n+k-2)} (\sigma_1 - 2\epsilon)^{k-1} e^{-\frac{2\epsilon}{\sigma_1} n}.
\end{aligned}$$

**(1-iii) Summary:** Therefore, we have the following results.

$$(\dagger_U) \leq \begin{cases} \frac{3(n-1)}{4} e^{-\frac{2\epsilon}{\sigma_1} n} & k < 1, \\ \frac{n \log(n\sigma_1)}{2} e^{-\frac{2\epsilon}{\sigma_1} n} & k = 1 \\ \frac{n^k(n-1)}{n+k-2} \frac{(\sigma_1 - 2\epsilon)^{k-1}}{2} e^{-\frac{2\epsilon}{\sigma_1} n} & k > 1. \end{cases} \tag{3.22}$$

**(2) Upper bound of  $(\diamond_U)$**

From  $\mathbb{1}[\bar{x}_1^{(n)} = x_1^{(1)} + 1/n] = \mathbb{1}[x_1^{(n)} \leq x_1^{(1)} + 1/n]$ , it holds that

$$\mathbb{1}[\bar{x}_1^{(n)} = x_1^{(1)} + 1/n] \left( \frac{2(\mu_1 - \epsilon - x_1^{(1)})}{\bar{x}_1^{(n)} - x_1^{(1)}} \right) = \mathbb{1}[x_1^{(n)} \leq x_1^{(1)} + 1/n] 2n(\mu_1 - \epsilon - x_1^{(1)}).$$

Therefore, applying Lemma 3.12, we obtain

$$\begin{aligned}
(\diamond_U) &= \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \int_y^{y + \frac{1}{n}} \frac{n(n-1)}{\sigma_1^n} (2n(\mu_1 - \epsilon - y))^{n+k-2} (z-y)^{n-2} dz dy \\
&= \int_{\mu_1 - \frac{\sigma_1}{2}}^{\mu_1 - \epsilon} \frac{n^k}{\sigma_1^n} (2(\mu_1 - \epsilon - y))^{n+k-2} dy \\
&= \frac{1}{2} \int_0^{\sigma_1 - 2\epsilon} \frac{n^k}{\sigma_1^n} w^{n+k-2} dw && \text{By a change of variables} \\
&= \frac{1}{2} \frac{n^k}{n+k-1} (\sigma_1 - 2\epsilon)^{k-1} \left( \frac{\sigma_1 - 2\epsilon}{\sigma_1} \right)^n \\
&\leq \frac{1}{2} \frac{n^k}{n+k-1} (\sigma_1 - 2\epsilon)^{k-1} e^{-\frac{2\epsilon}{\sigma_1} n}. \tag{3.23}
\end{aligned}$$

### (3) Conclusion

Therefore, by combining (3.19), (3.20), (3.22), and (3.23) with (3.18) and (3.17), we have for  $\epsilon > 0$  and

$$\psi(n; k) = \begin{cases} \mathcal{O}(n) & k \leq 0, \\ \mathcal{O}(n \log n) & k \in (0, 1], \\ \mathcal{O}(n^k) & k > 1 \end{cases}$$

that

$$\begin{aligned}
&\sum_{n=n_0}^T \mathbb{E} \left[ \frac{1 - p_n(\epsilon | \theta_{1,n})}{p_n(\epsilon | \theta_{1,n})} \right] \\
&\leq \sum_{n=n_0}^T 4e^{-\frac{\epsilon}{2\sigma} n} + e^{-\frac{\epsilon}{\sigma + \epsilon} (n-1)} + \psi(n; k) e^{-\frac{2\epsilon}{\sigma} n} + \mathcal{O}(n^{k-1}) e^{-\frac{2\epsilon}{\sigma} n} \\
&\leq \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-\max(2, k+1)}) + \mathcal{O}(\epsilon^{-\max(1, k)}) \\
&= \mathcal{O}(\epsilon^{-\max(2, k+1)}),
\end{aligned}$$

which concludes the proof of Lemma 3.9 for the case of TS-T.  $\square$

### 3.7.5 Proofs of technical lemmas for Lemma 3.8

In this section, we provide the detailed proofs of Lemmas 3.13 and 3.14. Notice that Lemma 3.14 is related to the main regret term of the policy.

*Proof of Lemma 3.13.* By the definition of  $x_i^{(1)}$  and  $x_i^{(n)}$ , which is the first order statistic and the last order statistic of  $X_{i,n}$ , respectively, we have

$$\begin{aligned}
\mathbb{P} \left[ x_i^{(1)} \geq \mu_i - \frac{\sigma_i}{2} + \epsilon \right] &= \mathbb{P} \left[ \forall s \in [n] : x_{i,s} \geq \mu_i - \frac{\sigma_i}{2} + \epsilon \right] \\
&= \left( 1 - \frac{\epsilon}{\sigma_i} \right)^n \leq \exp \left( -\frac{\epsilon}{\sigma_i} n \right).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathbb{P} \left[ \bar{x}_i^{(n)} \leq \mu_i + \frac{\sigma_i}{2} - \epsilon \right] &= \mathbb{P} \left[ \left\{ \forall s \in [n] : x_{i,s} \leq \mu_i + \frac{\sigma_i}{2} - \epsilon \right\}, x_i^{(1)} \leq \mu_i + \frac{\sigma_i}{2} - \epsilon - \frac{1}{n} \right] \\
&\leq \mathbb{P} \left[ \forall s \in [n] : x_{i,s} \leq \mu_i + \frac{\sigma_i}{2} - \epsilon \right] = \mathbb{P} \left[ x_i^{(n)} \leq \mu_i + \frac{\sigma_i}{2} - \epsilon \right] \\
&\leq \left( 1 - \frac{\epsilon}{\sigma_i} \right)^n \leq \exp \left( -\frac{\epsilon}{\sigma_i} \right),
\end{aligned}$$

which concludes the proof.  $\square$

*Proof of Lemma 3.14.* The overall proofs for both TS and TS-T are almost the same. For simplicity, we fix a time index  $t$  and denote  $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid T(X_{i,N_i(t)})] = \mathbb{P}[\cdot \mid \theta_{i,n}]$  and  $N_i(t) = n$  in this proof, where  $\theta_{i,n} = (x_i^{(1)}, x_i^{(n)})$

**Under TS** Under the condition  $\{\mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t)\}$ , by the law of total expectation, it holds that

$$\begin{aligned}
\mathbb{E}[\mathbb{1}[i(t) = i, \tilde{\mu}^*(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t) = n]] \\
\leq \mathbb{E}_{\theta_{i,n}} [\mathbb{P}_t[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon), N_i(t) = n]].
\end{aligned}$$

Since  $\tilde{\mu}_i | \tilde{\sigma}_i \sim \text{Uniform}_{ab} \left( x_i^{(n)} - \frac{\tilde{\sigma}_i}{2}, x_i^{(1)} + \frac{\tilde{\sigma}_i}{2} \right)$ , if  $x_i^{(n)} - \frac{\tilde{\sigma}_i}{2} \geq \mu_1 - \epsilon$  holds, then  $\tilde{\mu}_i \geq \mu_1 - \epsilon$  holds with probability 1. Therefore, we have

$$\begin{aligned}
\mathbb{P}_t[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] &= \mathbb{P}_t \left[ x_i^{(n)} - \frac{\tilde{\sigma}_i}{2} \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon) \right] \\
&\quad + \mathbb{P}_t \left[ \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, x_i^{(n)} - \frac{\tilde{\sigma}_i}{2} \leq \mu_1 - \epsilon \leq x_i^{(1)} + \frac{\tilde{\sigma}_i}{2}, \mathcal{E}_{i,n}(\epsilon) \right] \\
&\leq \mathbb{P}_t \left[ \frac{\tilde{\sigma}_i}{2} \leq \mu_i + \frac{\sigma_i}{2} - (\mu_1 - \epsilon), \mathcal{E}_{i,n}(\epsilon) \right] \\
&\quad + \mathbb{P}_t \left[ \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, x_i^{(n)} - \frac{\tilde{\sigma}_i}{2} \leq \mu_1 - \epsilon \leq x_i^{(1)} + \frac{\tilde{\sigma}_i}{2}, \mathcal{E}_{i,n}(\epsilon) \right].
\end{aligned}$$

Since  $\tilde{\sigma}_i \geq x_i^{(n)} - x_i^{(1)} = \hat{\sigma}_{i,n}$  always holds from the sampling procedure of TS, we have

$$\mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] \tilde{\sigma}_i(t) \geq \mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] \left( x_i^{(n)} - x_i^{(1)} \right) \geq \mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] (\sigma_i - 2\epsilon).$$

By the choice of  $\epsilon < \frac{\Delta_i}{2}$ , it holds that

$$\sigma_i - 2\epsilon \geq \sigma_i + 2\epsilon - 2\Delta_i = \sigma_i + 2\epsilon + 2(\mu_i - \mu_1),$$

which implies

$$\mathbb{P}_t \left[ \frac{\tilde{\sigma}_i}{2} \leq \mu_i + \frac{\sigma_i}{2} - (\mu_1 - \epsilon), \mathcal{E}_{i,n}(\epsilon) \right] = 0.$$

Then, it holds that

$$\begin{aligned}
\mathbb{P}_t \left[ \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, x_i^{(n)} - \frac{\tilde{\sigma}_i}{2} \leq \mu_1 - \epsilon \leq x_i^{(1)} + \frac{\tilde{\sigma}_i}{2}, \mathcal{E}_{i,n}(\epsilon) \right] \\
= \mathbb{P}_t \left[ \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mu_1 - \epsilon \leq x_i^{(1)} + \frac{\tilde{\sigma}_i}{2}, \mathcal{E}_{i,n}(\epsilon) \right] \\
\leq \mathbb{P}_t [\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \sigma_i + 2\Delta_i - 4\epsilon \leq \tilde{\sigma}_i, \mathcal{E}_{i,n}(\epsilon)],
\end{aligned}$$

where the inequality holds from  $x_i^{(1)} \leq \mu_i - \frac{\sigma_i}{2} + \epsilon$  on  $\mathcal{E}_{i,n}(\epsilon)$ . Therefore, by taking expectation, we have for a constant  $B_i := \sigma_i + 2\Delta_i - 4\epsilon$  that

$$\begin{aligned}
& \mathbb{E} \left[ \mathbb{P}_t[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] \right] \\
& \leq \int_{\sigma_i + 2\Delta_i - 4\epsilon}^{\infty} \pi^k(s|\theta_{i,n}) \int_{\mu_1 - \epsilon}^{\mu_i + \frac{s - \sigma_i}{2} + \epsilon} \pi^k(m|\theta_{i,n}, \tilde{\sigma}_i = s) dm ds \\
& = \int_{\sigma_i + 2\Delta_i - 4\epsilon}^{\infty} \frac{1}{s - (x_i^{(n)} - x_i^{(1)})} \left( \frac{s - \sigma_i}{2} - \Delta_i + 2\epsilon \right) \pi^k(s|\theta_{i,n}) ds \\
& = (x_i^{(n)} - x_i^{(1)})^{n+k-2} \int_{\sigma_i + 2\Delta_i - 4\epsilon}^{\infty} \frac{(n+k-1)(n+k-2)}{s^{n+k}} \cdot \left( \frac{s - \sigma_i}{2} - \Delta_i + 2\epsilon \right) ds \\
& = \frac{(x_i^{(n)} - x_i^{(1)})^{n+k-2}}{2} \int_{B_i}^{\infty} \frac{(n+k-1)(n+k-2)}{s^{n+k}} (s - B_i) ds \\
& = \frac{(x_i^{(n)} - x_i^{(1)})^{n+k-2}}{2} \left( \frac{n+k-1}{B_i^{n+k-2}} - \frac{n+k-2}{B_i^{n+k-2}} \right) \\
& \leq \frac{1}{2} \left( \frac{\sigma_i}{\sigma_i + 2\Delta_i - 4\epsilon} \right)^{n+k-2} = \frac{1}{2} \left( \frac{1}{1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}} \right)^{n+k-2}, \tag{3.24}
\end{aligned}$$

where the last inequality holds from  $x_i^{(1)} \geq \mu_i - \frac{\sigma_i}{2}$  and  $x_i^{(n)} \leq \mu_i + \frac{\sigma_i}{2}$ . For the arm  $i \neq 1$  and arbitrary  $n_i > n_0$ , we have

$$\begin{aligned}
& \sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = i, \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)]] \\
& \leq n_i + \sum_{t=n_0K+1}^T \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t) \geq n_i] \\
& \leq n_i + \sum_{t=n_0K+1}^T \frac{1}{2} \left( \frac{1}{1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}} \right)^{n_i+k-2} \\
& = n_i + \frac{T}{2} \left( \frac{1}{1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}} \right)^{n_i+k-2}.
\end{aligned}$$

Letting  $n_i = \max(2-k, 0) + \frac{\log T}{\log \left( 1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i} \right)}$  concludes the proof of Lemma 3.14 for the case of TS.

**Under TS-T** From the sampling rule of TS-T, it holds that

$$\begin{aligned}
& \mathbb{E}[\mathbb{1}[i(t) = i, \tilde{\mu}^*(t) \geq \mu_1 - \epsilon, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), N_i(t) = n]] \\
& \leq \mathbb{E}_{\theta_{i,n}} [\mathbb{P}_t[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \bar{\mathcal{E}}_{i,n}(\epsilon), N_i(t) = n]].
\end{aligned}$$

Therefore, the only differences from the proof of the case of TS are  $\bar{x}_i^{(n)}$  and  $\bar{\mathcal{E}}$  instead of  $x_i^{(n)}$  and  $\mathcal{E}$ , respectively. By following the same steps as under TS, we have an additional

restriction in (3.24), where the last inequality holds for TS-T when  $\frac{1}{n} \leq \sigma_i$  to satisfy  $\bar{x}_i^{(n)} \leq \mu_i + \frac{\sigma_i}{2}$ . Therefore, for arm  $i \neq 1$  and arbitrary  $n_i > \max\left(n_0, \frac{1}{\sigma_i}\right)$ , we have

$$\begin{aligned}
& \sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = i, \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \bar{\mathcal{E}}_{i,n}(\epsilon)]] \\
& \leq n_i + \sum_{t=n_0K+1}^T \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), N_i(t) \geq n_i] \\
& \leq n_i + \sum_{t=n_0K+1}^T \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t) \geq n_i] \\
& \leq n_i + \sum_{t=n_0K+1}^T \frac{1}{2} \left( \frac{1}{1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}} \right)^{n_i + k - 2} \\
& = n_i + \frac{T}{2} \left( \frac{1}{1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}} \right)^{n_i + k - 2}.
\end{aligned}$$

Letting  $n_i = \max\left(\frac{1}{\sigma_i}, 2 - k\right) + \frac{\log T}{\log\left(1 + \frac{2\Delta_i - 4\epsilon}{\sigma_i}\right)}$  concludes the proof.  $\square$

### 3.7.6 Proof of Lemma 3.10

The proof of Lemma 3.10 can be easily derived from the lemmas below, which are the counterparts of Lemmas 3.13 and 3.14 in the Gaussian bandits.

Note that the regret lower bound in (3.6) is invariant under the location and scale transformation, which implies that

$$\begin{aligned}
& \inf_{(\mu, \sigma): \mu > \mu_1} \text{KL}(\text{Gaussian}(\mu_i, \sigma_i); \text{Gaussian}(\mu, \sigma)) \\
& = \inf_{(\mu, \sigma): \mu > \mu_1} \text{KL}\left(\text{Gaussian}\left(\frac{\mu_i - a}{b}, \frac{\sigma_i}{b}\right); \text{Gaussian}\left(\frac{\mu - a}{b}, \frac{\sigma}{b}\right)\right).
\end{aligned}$$

In the remaining sections of this chapter, we consider the Gaussian bandit instance where  $(\mu_1, \sigma_1) = (0, 1)$  for simplicity since one can recover the original instance by the location and scale transformation. Similarly to the uniform bandits, let us define two events for  $n \in \mathbb{N}$  and  $i \in [K]$  that

$$\begin{aligned}
\mathcal{M}_\epsilon(t) &= \{\tilde{\mu}^*(t) \geq -\epsilon\}, \\
\mathcal{E}_{i,n}(\epsilon) &= \{\hat{x}_{i,n} \leq \mu_i + \epsilon, S_{i,n} \leq n(\sigma_i^2 + \epsilon)\}.
\end{aligned}$$

In this section,  $\theta_{i,n}$  denotes  $(\hat{x}_{i,n}, S_{i,n})$ , which are the sufficient statistic in the Gaussian models. To begin the proof, we first provide some known results in the Gaussian bandits.

**Lemma 3.15** (Lemma 9 in Honda and Takemura [2014]). *For any  $i \neq 1$ ,*

$$\mathbb{E} \left[ \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) = i, \mathcal{E}_{i,n}^c(\epsilon)] \right] \leq \mathcal{O}(\epsilon^{-2}).$$

**Lemma 3.16** (Lemma 4 in Honda and Takemura [2014]). *If  $\mu > \hat{x}_{i,n}$  and  $n \geq n_0$ , then*

$$\mathbb{P}[\tilde{\mu}_i \geq \mu | \hat{x}_{i,n}, S_{i,n}] \geq A_{n,k} \left( 1 + \frac{n(\mu - \hat{x}_{i,n})^2}{S_{i,n}} \right)^{-\frac{n+k-2}{2}} \quad (3.25)$$

and

$$\mathbb{P}[\tilde{\mu}_i \geq \mu | \hat{x}_{i,n}, S_{i,n}] \leq \frac{\sqrt{S_{i,n}}}{\mu - \hat{x}_{i,n}} \left( 1 + \frac{n(\mu - \hat{x}_{i,n})^2}{S_{i,n}} \right)^{-\frac{n+k-3}{2}}, \quad (3.26)$$

where

$$A_{n,k} = \frac{1}{2e^{1/6} \sqrt{\frac{n+k-1}{2}\pi}}.$$

*Proof of Lemma 3.10.* Let us first define an event on the truncated statistic

$$\bar{\mathcal{E}}_{i,n}(\epsilon) := \{ \hat{x}_{i,n} \leq \mu_i + \epsilon, \bar{S}_{i,n} \leq n(\sigma_i^2 + \epsilon) \}.$$

Similarly to the analysis of TS-T in the uniform bandits, we can decompose (GO) as

$$(\text{GO}) \leq \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1} [i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] + \Delta_i \mathbb{1} [i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}^c(\epsilon)].$$

From the definition of  $\bar{S}_{i,n} = \max(1, S_{i,n})$ , it holds that

$$\mathcal{E}_{i,n}(\epsilon) \subset \bar{\mathcal{E}}_{i,n}(\epsilon).$$

Therefore, from Lemma 3.15, we have

$$\begin{aligned} \mathbb{E}[(\text{GO})] &\leq \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{E} [i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] \\ &\quad + \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{E} [i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}^c(\epsilon)] \\ &\leq \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{E} [i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] + \mathcal{O}(\epsilon^{-2}). \end{aligned} \quad (3.27)$$

It remains to show the upper bound of the first term of (3.27). Let  $n_i > \frac{1}{\sigma_i^2}$  be arbitrary, where  $\mathcal{E}_{i,n_i}(\epsilon) = \bar{\mathcal{E}}_{i,n_i}(\epsilon)$  holds for any  $\epsilon > 0$ . Then, by injecting (3.26) in Lemma 3.16

$$\begin{aligned} &\mathbb{E} \left[ \sum_{t=Kn_0+1}^T \mathbb{1} \left[ i(t) = i, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t) \right] \right] \\ &\leq n_i + \sum_{t=Kn_0+1}^T \mathbb{P} [\tilde{\mu}_i(t) \geq -\epsilon, \bar{\mathcal{E}}_{i,N_i(t)}(\epsilon), N_i(t) \geq n_i] \\ &= n_i + \sum_{t=Kn_0+1}^T \mathbb{P} [\tilde{\mu}_i(t) \geq -\epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t) \geq n_i] \\ &\leq n_i + \sum_{t=Kn_0+1}^T \frac{\sqrt{\sigma_i^2 + \epsilon}}{\Delta_i - 2\epsilon} \left( 1 + \frac{(\Delta_i - 2\epsilon)^2}{\sigma_i^2 + \epsilon} \right)^{-\frac{n+k-3}{2}} \\ &= n_i + T \frac{\sqrt{\sigma_i^2 + \epsilon}}{\Delta_i - 2\epsilon} \exp \left( -(n+k-3) \frac{1}{2} \log \left( 1 + \frac{(\Delta_i - 2\epsilon)^2}{\sigma_i^2 + \epsilon} \right) \right). \end{aligned}$$

Letting  $n_i = \max \left( \sigma_i^{-2}, \frac{\log T}{\frac{1}{2} \log \left( 1 + \frac{(\Delta_i - 2\epsilon)^2}{\sigma_i^2 + \epsilon} \right)} + 3 - k \right)$  completes the proof.  $\square$

### 3.7.7 Proof of Lemma 3.11

Firstly, we introduce some technical results from Honda and Takemura [2014] before beginning the proof.

**Lemma 3.17** (Some results in Honda and Takemura [2014]). *For  $n \geq n_0$  and  $\epsilon > 0$ , it holds that*

$$\begin{aligned}\mathbb{P}[-\epsilon \leq \hat{\mu}_{1,n} \leq -\epsilon/2] &\leq e^{-\frac{\epsilon^2}{8}n}, \\ \mathbb{P}[-\epsilon/2 \leq \hat{\mu}_{1,n}, S_{1,n} \geq 2n] &\leq e^{-\frac{1-\log 2}{2}n}.\end{aligned}$$

**Lemma 3.18** (Lemma 10 of Honda and Takemura [2014]). *For  $z \geq 1/2$*

$$e^{-2/3} \leq \frac{\Gamma(z + \frac{1}{2})}{\Gamma(z)} \leq e^{1/6} \sqrt{z}.$$

Next, we introduce two functions and their corresponding integral representations to analyze the term induced by TS-T.

**Definition 3.19.** The confluent hypergeometric function of the second kind  $U(a, b, z)$ , a.k.a. Tricomi's function [Tricomi, 1947], is a solution of Kummer's equation

$$z \frac{d^2 w}{dz^2} + (b - z) \frac{dw}{dz} - aw = 0,$$

which can be uniquely determined by satisfying for arbitrary small constant  $\epsilon > 0$

$$U(a, b, z) \sim z^{-a}, \quad z \rightarrow \infty, |\text{ph}z| \leq \frac{3}{2}\pi - \epsilon.$$

Here,  $\text{ph}z$  denotes the phase of  $z \in \mathbb{C}$ . It has its integral representation for  $a, b \in \mathbb{R}_+$  such that  $b > a$  and  $z \in \mathbb{R}_+$  as follows [Olver et al., 2010, 13.4.4]:

$$U(a, b, z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{b-a-1} dt. \quad (3.28)$$

**Definition 3.20.** The modified Bessel function of the second kind is a standard solution of the modified Bessel's equation

$$z^2 \frac{d^2 w}{dz^2} + z \frac{dw}{dz} - (z^2 + v^2)w = 0,$$

which can be uniquely determined by satisfying that

$$K_v(z) \sim \sqrt{\frac{\pi}{2z}} e^{-z}, \quad z \rightarrow \infty, |\text{ph}z| < \frac{3}{2}\pi.$$

It has the integral representation as follows [Olver et al., 2010, 10.32.9]:

$$K_v(z) = \int_0^\infty e^{-z \cosh t} \cosh(vt) dt, \quad (3.29)$$

where  $K_v(z) = K_{-v}(z)$  holds.

Then, we provide two technical lemmas, whose proofs are given in Section 3.7.8.



**Lemma 3.21.** Let  $\Gamma(s, x) = \int_x^\infty t^{s-1} e^{-t} dt$  denote the upper incomplete gamma function. Then,

$$\begin{aligned} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-\frac{k+1}{2}} \Gamma\left(\frac{n}{2}, \frac{1}{2(1-w)}\right) dw &\leq \Gamma\left(\frac{n}{2}\right) \int_0^1 w^{-\frac{1}{2}} (1-w)^{-1+\frac{2}{n\frac{k+1}{2}}} dw \\ &= \Gamma\left(\frac{n}{2}\right) B\left(\frac{1}{2}, \frac{2}{n\frac{k+1}{2}}\right) \end{aligned}$$

is valid for  $k \in \{1, 2\}$  and  $n \geq n_0 = \max(2, 4-k)$ , where  $B(z_1, z_2)$  denotes the Beta function.

**Lemma 3.22.** For  $a, b, z \in \mathbb{R}$ , let  $U(a, b, z)$  denote the confluent hypergeometric function of the second kind. Then,

$$U\left(\frac{1}{2}, b, \frac{1}{2}\right) \leq \frac{2^b}{\Gamma\left(\frac{1}{2}\right)} \Gamma\left(b - \frac{1}{2}\right)$$

is valid for  $b \in \{\frac{m}{2} : m \in \mathbb{Z}_{\geq 4}\}$ .

Finally, we provide the numerical results of the computation of the modified Bessel function of the second kind, which is used several times in the proof.

**Fact 3.23** (Table 2 in Watson [1922]). Let  $K_\nu(z)$  denote the modified Bessel function of the second kind. Then, the followings are the results of numerical computations evaluated to 6S.

$$\begin{aligned} e^{0.24} K_0(0.24) &= 2.00835 \\ e^{0.24} K_1(0.24) &= 4.98213 \end{aligned}$$

*Proof of Lemma 3.11.* Let us define  $\bar{\theta}_{1,n} = (\hat{\mu}_{1,n}, \bar{S}_{1,n})$ . Similarly to Lemma 3.9 in the uniform model, let us consider the following decomposition:

$$\begin{aligned} \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{M}_\epsilon^c(t)] &= \sum_{n=n_0}^T \sum_{t=Kn_0+1}^T \mathbb{1}\left[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n\right] \\ &= \sum_{n=n_0}^T \sum_{m=1}^T \mathbb{1}\left[m \leq \sum_{t=Kn_0+1}^T \mathbb{1}\left[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n\right]\right] \\ &\leq \mathbb{E} \left[ \sum_{n=n_0}^T \sum_{m=1}^T (1 - p_n(\epsilon|\bar{\theta}_{1,n}))^m \right] \\ &\leq \sum_{n=n_0}^T \mathbb{E} \left[ \frac{1 - p_n(\epsilon|\bar{\theta}_{1,n})}{p_n(\epsilon|\bar{\theta}_{1,n})} \right], \end{aligned}$$

where  $p_n(\epsilon|\bar{\theta}_{1,n}) = \mathbb{P}[\tilde{\mu}_1 \geq -\epsilon|\hat{\mu}_{1,n}, \bar{S}_{1,n}]$ . Since the Student's  $t$ -distribution is symmetric about its location parameter,  $\mathbb{1}[\hat{\mu}_{1,n} \geq -\epsilon] p_n(\epsilon|\bar{\theta}) \geq 1/2$  holds. Therefore, we have

$$\mathbb{E} \left[ \frac{1 - p_n(\epsilon|\bar{\theta}_{1,n})}{p_n(\epsilon|\bar{\theta}_{1,n})} \right] \leq 2 \mathbb{E} [\mathbb{1}[\hat{\mu}_{1,n} \geq -\epsilon] (1 - p_n(\epsilon|\bar{\theta}_{1,n}))] + \mathbb{E} \left[ \frac{\mathbb{1}[\hat{\mu}_{1,n} \leq -\epsilon]}{p_n(\epsilon|\bar{\theta}_{1,n})} \right]. \quad (3.30)$$

By applying Lemma 3.17 to the first term in (3.30), it holds that

$$\begin{aligned}
& \mathbb{E} \left[ \mathbb{1}[\hat{\mu}_{1,n} \geq -\epsilon](1 - p_n(\epsilon|\bar{\theta}_{1,n})) \right] \\
&= \mathbb{P}[-\epsilon \leq \hat{\mu}_{1,n} \leq -\epsilon/2] + \mathbb{P}[-\epsilon/2 \leq \hat{\mu}_{1,n}, \bar{S}_{1,n} \geq 2n] \\
&\quad + \mathbb{E} \left[ \mathbb{1}[-\epsilon/2 \leq \hat{\mu}_{1,n}, \bar{S}_{1,n} \leq 2n](1 - p_n(\epsilon|\bar{\theta}_{1,n})) \right] \\
&= \mathbb{P}[-\epsilon \leq \hat{\mu}_{1,n} \leq -\epsilon/2] + \mathbb{P}[-\epsilon/2 \leq \hat{\mu}_{1,n}, S_{1,n} \geq 2n] \\
&\quad + \mathbb{E} \left[ \mathbb{1}[-\epsilon/2 \leq \hat{\mu}_{1,n}, \bar{S}_{1,n} \leq 2n](1 - p_n(\epsilon|\bar{\theta}_{1,n})) \right] \\
&\leq e^{-\frac{\epsilon^2}{8}n} + e^{-\frac{1-\log 2}{2}n} + \mathbb{E} \left[ \mathbb{1}[-\epsilon/2 \leq \hat{\mu}_{1,n}, \bar{S}_{1,n} \leq 2n](1 - p_n(\epsilon|\bar{\theta}_{1,n})) \right],
\end{aligned}$$

where the second equality holds from the definition of  $\bar{S}_{1,n} = \max(1, S_{1,n})$ , which implies  $\{\bar{S}_{1,n} \geq 2n\} = \{S_{1,n} \geq 2n\}$  for any  $n \in \mathbb{N}$ . From the symmetry of  $t$ -distribution, it holds that

$$\begin{aligned}
1 - p_n(\epsilon|\bar{\theta}_{1,n}) &= \int_{-\infty}^{-\epsilon} f_{n+k-2}^t(y; \hat{x}_{1,n}, \bar{S}_{1,n}) dy = \int_{\epsilon}^{\infty} f_{n+k-2}^t(y; -\hat{x}_{1,n}, \bar{S}_{1,n}) dy \\
&= \int_{2\hat{x}_{1,n}+\epsilon}^{\infty} f_{n+k-2}^t(y; \hat{x}_{1,n}, \bar{S}_{1,n}) dy \\
&= \mathbb{P}[\hat{\mu}_1 \geq 2\hat{x}_{i,n} + \epsilon | \hat{x}_{i,n}, \bar{S}_{i,n}].
\end{aligned}$$

From (3.26) in Lemma 3.16, it holds that

$$\mathbb{E} \left[ \mathbb{1}[-\epsilon/2 \leq \hat{\mu}_{1,n}, \bar{S}_{1,n} \leq 2n](1 - p_n(\epsilon|\bar{\theta}_{1,n})) \right] \leq \frac{2\sqrt{2}}{\epsilon} \left( 1 + \frac{\epsilon^2}{8} \right)^{-\frac{n+k-3}{2}}.$$

Therefore, the first term in (3.30) can be bounded as

$$2\mathbb{E} \left[ \mathbb{1}[\hat{\mu}_{1,n} \geq -\epsilon](1 - p_n(\epsilon|\bar{\theta}_{1,n})) \right] \leq 2e^{-\frac{\epsilon^2}{8}n} + 2e^{-\frac{1-\log 2}{2}n} + \frac{4\sqrt{2}}{\epsilon} \left( 1 + \frac{\epsilon^2}{8} \right)^{-\frac{n+k-3}{2}}. \quad (3.31)$$

Note that the last term in (3.30) was a problematic term for TS with priors  $k \geq 1$  [Honda and Takemura, 2014]. However, we showed that such a problem could be resolved by replacing  $S_{1,n}$  with  $\bar{S}_{1,n}$ .

Finally, we evaluate the last term in (3.30). From the definition of  $\bar{S}_{1,n}$ , it holds that  $\mathbb{1}[\bar{S}_{1,n} > 1] = \mathbb{1}[\bar{S}_{1,n} = S_{1,n}]$  and  $\mathbb{1}[\bar{S}_{1,n} = 1] = \mathbb{1}[S_{1,n} \leq 1]$ . Therefore,

$$\begin{aligned}
\mathbb{E} \left[ \frac{\mathbb{1}[\hat{\mu}_{1,n} \leq -\epsilon]}{p_n(\epsilon|\bar{\theta}_{1,n})} \right] &= \mathbb{E} \left[ \frac{\mathbb{1}[\hat{\mu}_{1,n} \leq -\epsilon, \bar{S}_{1,n} > 1]}{p_n(\epsilon|\bar{\theta}_{1,n})} \right] + \mathbb{E} \left[ \frac{\mathbb{1}[\hat{\mu}_{1,n} \leq -\epsilon, \bar{S}_{1,n} = 1]}{p_n(\epsilon|\bar{\theta}_{1,n})} \right] \\
&= \underbrace{\mathbb{E} \left[ \frac{\mathbb{1}[\hat{\mu}_{1,n} \leq -\epsilon, S_{1,n} > 1]}{p_n(\epsilon|\bar{\theta}_{1,n})} \right]}_{(\dagger_G)} + \underbrace{\mathbb{E} \left[ \frac{\mathbb{1}[\hat{\mu}_{1,n} \leq -\epsilon, S_{1,n} \leq 1]}{p_n(\epsilon|\bar{\theta}_{1,n})} \right]}_{(\diamond_G)}. \quad (3.32)
\end{aligned}$$

From the sampling distributions of  $\hat{x}_{i,n}$  and  $S_{i,n}$  in (3.7) and (3.25) in Lemma 3.16, we obtain

$$(\dagger_G) = \frac{1}{A_{n,k}} \int_{-\infty}^{-\epsilon} \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}} \int_1^{\infty} \left( 1 + \frac{n(x+\epsilon)^2}{s} \right)^{\frac{n+k-2}{2}} \frac{s^{\frac{n-3}{2}} e^{-\frac{s}{2}}}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} ds dx$$

and

$$(\diamond_G) = \frac{1}{A_{n,k}} \mathbb{P}[S_{1,n} \leq 1] \int_{-\infty}^{-\epsilon} \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}} (1 + n(x+\epsilon)^2)^{\frac{n+k-2}{2}} dx$$

### Upper bound of $(\dagger_G)$

For  $k < 1$ , Lemma 8 in Honda and Takemura [2014] showed that

$$\frac{1}{A_{n,k}} \int_{-\infty}^{-\epsilon} \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}} \int_0^\infty \left(1 + \frac{n(x+\epsilon)^2}{s}\right)^{\frac{n+k-2}{2}} \frac{s^{\frac{n-3}{2}} e^{-\frac{s}{2}}}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} ds dx \leq \mathcal{O}(ne^{-n\epsilon^2}).$$

Therefore, the following result immediately follows for  $k < 1$ :

$$(\dagger_G) \leq \mathcal{O}(ne^{-n\epsilon^2}).$$

In the remaining proof, we focus on the case of  $k = 1, 2$ , which corresponds to the reference prior and the Jeffreys prior, respectively. Since  $x^2 \geq (x + \epsilon)^2 + \epsilon^2$  holds for  $x \leq -\epsilon$ , it holds that

$$(\dagger_G) \leq \frac{e^{-\frac{n\epsilon^2}{2}}}{A_{n,k}} \int_{-\infty}^{-\epsilon} \sqrt{\frac{n}{2\pi}} e^{-\frac{n(x+\epsilon)^2}{2}} \int_1^\infty \left(1 + \frac{n(x+\epsilon)^2}{s}\right)^{\frac{n+k-2}{2}} \frac{s^{\frac{n-3}{2}} e^{-\frac{s}{2}}}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} ds dx \quad (3.33)$$

Let us consider the change of variables

$$(x, s) = \left(-\epsilon - \sqrt{\frac{2zw}{n}}, 2z(1-w)\right),$$

which gives

$$dx ds = \sqrt{\frac{2z}{nw}} dz dw.$$

Then we obtain for  $k \leq 2$

$$\begin{aligned} (\dagger_G) &\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{A_{n,k}} \int_0^1 \int_{\frac{1}{2(1-w)}}^\infty \left(1 + \frac{w}{1-w}\right)^{\frac{n+k-2}{2}} \\ &\quad \cdot \sqrt{\frac{n}{2\pi}} e^{-zw} \frac{(z(1-w))^{\frac{n-3}{2}} e^{-z(1-w)}}{2\Gamma\left(\frac{n-1}{2}\right)} \sqrt{\frac{2z}{nw}} dz dw \\ &= \frac{e^{-\frac{n\epsilon^2}{2}}}{2\sqrt{\pi} A_{n,k} \Gamma\left(\frac{n-1}{2}\right)} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-\frac{k+1}{2}} \int_{\frac{1}{2(1-w)}}^\infty e^{-z} z^{\frac{n}{2}-1} dz dw \\ &= \frac{e^{-\frac{n\epsilon^2}{2}}}{2\sqrt{\pi} A_{n,k} \Gamma\left(\frac{n-1}{2}\right)} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-\frac{k+1}{2}} \Gamma\left(\frac{n}{2}, \frac{1}{2(1-w)}\right) dw \quad (3.34) \\ &\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{2\sqrt{\pi} A_{n,k} \Gamma\left(\frac{n-1}{2}\right)} \Gamma\left(\frac{n}{2}\right) B\left(\frac{1}{2}, \frac{2}{n^{\frac{k+1}{2}}}\right) \quad \text{by Lemma 3.21} \\ &\leq 2ne^{-\frac{\epsilon^2}{2}n} B\left(\frac{1}{2}, \frac{2}{n^{\frac{k+1}{2}}}\right) \quad \text{by Lemma 3.18} \\ &= 2ne^{-\frac{\epsilon^2}{2}n} \frac{\Gamma(1/2)\Gamma\left(\frac{2}{n^{\frac{k+1}{2}}}\right)}{\Gamma\left(\frac{1}{2} + \frac{2}{n^{\frac{k+1}{2}}}\right)} \leq 2ne^{-\frac{\epsilon^2}{2}n} \sqrt{\pi} \Gamma\left(\frac{2}{n^{\frac{k+1}{2}}}\right). \quad \text{by (3.35)} \end{aligned}$$

where we used

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \quad \Gamma(x) \geq 1, \quad \text{for } x \in (0, 1). \quad (3.35)$$

By the Laurent expansion of the Gamma function around  $z = 0$ , it holds that

$$\Gamma(z) = \frac{1}{z} - \gamma + \frac{1}{2} \left( \gamma^2 + \frac{\pi}{6} \right) z - \mathcal{O}(z^2),$$

where  $\gamma$  denotes the Euler–Mascheroni constant, such that  $\gamma \in (0.57, 0.58)$ .

Then, for  $k \geq 1$  and  $n \geq 2$ , it holds that

$$\begin{aligned} \Gamma\left(\frac{2}{n^{\frac{k+1}{2}}}\right) &\leq \frac{1}{2} n^{\frac{k+1}{2}} - \gamma + \frac{1}{2} \left( \gamma^2 + \frac{\pi}{6} \right) \frac{2}{n^{\frac{k+1}{2}}} \\ &\leq \frac{1}{2} n^{\frac{k+1}{2}} - \gamma + \frac{1}{2} \left( \gamma^2 + \frac{\pi}{6} \right) \frac{2}{n} \\ &\leq \frac{1}{2} n^{\frac{k+1}{2}}. \end{aligned}$$

Therefore, for  $k \in \{1, 2\}$ , it holds that

$$(\dagger_G) \leq \mathcal{O}(n^{\frac{k+3}{2}} e^{-n\epsilon^2}).$$

Note that for  $k \in (1, 2)$ , the integral in (3.34) is increasing function with respect to  $k \in [1, 2]$ , which gives for  $k \in [1, 2]$  that

$$(\dagger_G) \leq \mathcal{O}\left(n^{\frac{5}{2}} e^{-n\epsilon^2}\right).$$

Therefore, we have

$$(\dagger_G) \leq \begin{cases} \mathcal{O}\left(n e^{-n\epsilon^2}\right) & \text{if } k < 1, \\ \mathcal{O}\left(n^{\frac{\lceil k \rceil + 3}{2}} e^{-n\epsilon^2}\right) & \text{if } k \in [1, 2], \end{cases} \quad (3.36)$$

where  $\lceil \cdot \rceil$  denotes the ceiling function.

### Upper bound of $(\diamond_G)$

Similarly to (3.33), it holds that

$$\begin{aligned} (\diamond_G) &\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{A_{n,k}} \mathbb{P}[S_{1,n} \leq 1] \int_{-\infty}^{-\epsilon} \sqrt{\frac{n}{2\pi}} e^{-\frac{n(x+\epsilon)^2}{2}} (1 + n(x+\epsilon)^2)^{\frac{n+k-2}{2}} dx \\ &= \frac{e^{-\frac{n\epsilon^2}{2}}}{A_{n,k}} \sqrt{\frac{n}{2\pi}} \mathbb{P}[S_{1,n} \leq 1] \int_{-\infty}^0 e^{-\frac{nx^2}{2}} (1 + nx^2)^{\frac{n+k-2}{2}} dx \\ &= \frac{e^{-\frac{n\epsilon^2}{2}}}{A_{n,k}} \sqrt{\frac{n}{2\pi}} \mathbb{P}[S_{1,n} \leq 1] \int_0^{\infty} e^{-\frac{nx^2}{2}} (1 + nx^2)^{\frac{n+k-2}{2}} dx. \end{aligned}$$

Here, Recall the integral representation of the confluent hypergeometric function of the second kind in (3.28), which is

$$U(a, b, z) = \frac{1}{\Gamma(a)} \int_0^{\infty} e^{-zt} t^{a-1} (1+t)^{b-a-1} dt.$$

Therefore, by letting  $t = nx^2$ , we have

$$\begin{aligned} (\diamond_G) &\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{A_{n,k}} \sqrt{\frac{n}{2\pi}} \sqrt{\frac{1}{2n}} \mathbb{P}[S_{1,n} \leq 1] \int_0^{\infty} e^{-\frac{t}{2}} t^{-\frac{1}{2}} (1+t)^{\frac{n+k-2}{2}} dt \\ &\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{2A_{n,k}} \sqrt{\frac{1}{\pi}} \mathbb{P}[S_{1,n} \leq 1] U\left(\frac{1}{2}, \frac{n+k+1}{2}, \frac{1}{2}\right). \end{aligned}$$

From Lemma 3.22, we obtain

$$\begin{aligned}
(\diamond_G) &\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{2A_{n,k}} \sqrt{\frac{1}{\pi}} \mathbb{P}[S_{1,n} \leq 1] \left( \frac{3 \cdot 2^{\frac{n+k-1}{2}}}{\Gamma(\frac{1}{2})} \right) \Gamma\left(\frac{n+k}{2}\right) \\
&= \frac{3e^{-\frac{n\epsilon^2}{2}}}{\pi A_{n,k}} \mathbb{P}[S_{1,n} \leq 1] 2^{\frac{n+k-3}{2}} \Gamma\left(\frac{n+k}{2}\right). \tag{3.37}
\end{aligned}$$

For a random variable following the chi-squared distribution with the degree of freedom  $n$ , it holds for  $x \in (0, 1)$  that

$$\mathbb{P}[X \leq nx] \leq e^{-n \frac{x-1-\log x}{2}}.$$

Since  $S_{1,n} \sim \chi_{n-1}^2$  (recall that we consider the case  $\sigma_1 = 1$ ), by letting  $x = \frac{1}{n-1}$ , we obtain

$$\mathbb{P}[S_{1,n} \leq 1] \leq e^{-\frac{2-n+(n-1)\log(n-1)}{2}} = (n-1)^{-\frac{n-1}{2}} e^{\frac{n}{2}-1}. \tag{3.38}$$

By combining (3.38) with (3.37), we have for  $n \geq n_0 = 3$

$$(\diamond_G) \leq \frac{3e^{-\frac{n\epsilon^2}{2}}}{\pi A_{n,k}} (n-1)^{-\frac{n-1}{2}} e^{\frac{n}{2}-1} 2^{\frac{n+k-3}{2}} \Gamma\left(\frac{n+k}{2}\right).$$

From Stirling's formula,

$$\Gamma(z) \leq \sqrt{2\pi} e^{1/6} z^{z-\frac{1}{2}} e^{-z},$$

we have

$$\begin{aligned}
(\diamond_G) &\leq \frac{3e^{-\frac{n\epsilon^2}{2}}}{\pi A_{n,k}} (n-1)^{-\frac{n-1}{2}} e^{\frac{n}{2}-1} 2^{\frac{n+k-3}{2}} \sqrt{2\pi} e^{1/6} \left(\frac{n+k}{2}\right)^{\frac{n+k-1}{2}} e^{-\frac{n+k}{2}} \\
&= \frac{3e^{-\frac{n\epsilon^2}{2}}}{\sqrt{2\pi} A_{n,k}} (n-1)^{-\frac{n-1}{2}} e^{\frac{n}{2}-1} e^{1/6} (n+k)^{\frac{n+k-1}{2}} e^{-\frac{n+k}{2}} \\
&\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{\sqrt{2\pi} A_{n,k}} e^{1/6} e^{-\frac{k}{2}} (n-1)^{\frac{k}{2}} \left(\frac{n+k}{n-1}\right)^{\frac{n+k-1}{2}} \\
&\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{\sqrt{2\pi} A_{n,k}} e^{1/6} e^{-\frac{k}{2}} n^{\frac{k}{2}} \left(1 + \frac{k+1}{n-1}\right)^{\frac{n+k-1}{2}} \\
&\leq \frac{e^{-\frac{n\epsilon^2}{2}}}{\sqrt{2\pi} A_{n,k}} e^{5/6} n^{\frac{k}{2}} e^{\frac{k(k+1)}{2(n-1)}} \\
&= \mathcal{O}\left(n^{\frac{k+1}{2}} e^{-n\epsilon^2}\right). \tag{3.39}
\end{aligned}$$

## Conclusion

Therefore, by combining (3.36) and (3.39) with (3.32), we have for  $k \in \mathbb{Z}_{\leq 2}$

$$\mathbb{E} \left[ \frac{\mathbb{1}[\hat{x}_{1,n} \leq -\epsilon]}{p_n(\epsilon|\theta_{1,n})} \right] \leq \mathcal{O}\left(n^{\frac{m'}{2}} e^{-n\epsilon^2}\right) + \mathcal{O}\left(n^{\frac{\max(0,k+1)}{2}} e^{-n\epsilon^2}\right), \tag{3.40}$$

where  $m' = 2 \cdot \mathbb{1}[k \in \mathbb{Z}_{<1}] + (\lceil k \rceil + 3) \mathbb{1}[k \in [1, 2]]$ . Therefore, by injecting (3.31) and (3.40) to (3.30), we obtain for  $k \in \mathbb{Z}_{\leq 2}$

$$\begin{aligned}
(\text{BO}) &\leq \sum_{n=n_0}^T 2e^{-\frac{\epsilon^2}{8}n} + 2e^{-\frac{1-\log 2}{2}n} + \frac{4\sqrt{2}}{\epsilon} \left(1 + \frac{\epsilon^2}{8}\right)^{-\frac{n+k-3}{2}} \\
&\quad + \mathcal{O}\left(n^{\frac{m'}{2}} e^{-n\epsilon^2}\right) + \mathcal{O}\left(n^{\frac{\max(0,k+1)}{2}} e^{-n\epsilon^2}\right) \\
&\leq \mathcal{O}(\epsilon^{-2}) + \mathcal{O}(1) + \mathcal{O}(\epsilon^{-3}) + \mathcal{O}(\epsilon^{-(m'+2)}) + \mathcal{O}(\epsilon^{-(k+3)}).
\end{aligned}$$

Letting  $m = m' + 2 = 4 + \lceil k \rceil \mathbb{1}[k \in [1, 2]]$  concludes the proof.  $\square$

### 3.7.8 Proofs of technical lemmas for Lemma 3.11

In this section, we provide the all proofs of Lemmas 3.21 and 3.22 based on the mathematical induction.

*Proof of Lemma 3.21.* Define

$$g_k(n) := \int_0^1 w^{-\frac{1}{2}}(1-w)^{-\frac{k+1}{2}} \Gamma\left(\frac{n}{2}, \frac{1}{2(1-w)}\right) dw.$$

Here, we apply mathematical induction separately for both odd and even values of  $n$  for each  $k \in \{1, 2\}$ . We expect that one can extend the analysis for the case of  $k = 1$  to the general  $k$  by changing the parameter  $b$  of the hypergeometric function of the second kind  $U(a, b, z)$ .

**For the reference prior ( $k = 1$ )**

Let us consider the case of the even number  $n = 2m$ .

**(1) Even number** Since  $n_0 = \max(2, 4 - k)$ , it is sufficient to consider  $m \geq 2$  if  $k = 1$ .

**(1-i) Base case  $n = 4$**  From the definition of the upper incomplete gamma function, it holds that

$$\Gamma(2, x) = e^{-x}(x + 1).$$

By letting  $t = \frac{w}{1-w}$ , we have

$$\begin{aligned} g_1(4) &= \frac{1}{2\sqrt{e}} \int_0^\infty \sqrt{\frac{t+1}{t}} e^{-\frac{t}{2}} dt + \frac{1}{\sqrt{e}} \int_0^\infty \sqrt{\frac{1}{t(t+1)}} e^{-\frac{t}{2}} dt \\ &= \sqrt{\frac{\pi}{e}} \left( \frac{1}{2} U\left(\frac{1}{2}, 2, \frac{1}{2}\right) + U\left(\frac{1}{2}, 1, \frac{1}{2}\right) \right). \end{aligned}$$

Here, it holds as follows [Olver et al., 2010, 13.3.9 and 13.3.10]:

$$\begin{aligned} U(a, b, z) - aU(a+1, b, z) - U(a, b-1, z) &= 0, \\ (b-a)U(a, b, z) + U(a-1, b, z) - zU(a, b+1, z) &= 0, \end{aligned}$$

which gives

$$U\left(\frac{1}{2}, 2, \frac{1}{2}\right) = \frac{1}{4}U\left(\frac{3}{2}, 3, \frac{1}{2}\right) + \frac{1}{2}U\left(\frac{1}{2}, 1, \frac{1}{2}\right). \quad (3.41)$$

Let  $K_v(z)$  denote the modified Bessel function of the second kind defined in Definition 3.20. Then, we have the result in Olver et al. [2010, 13.6.10.] that

$$U\left(v + \frac{1}{2}, 2v + 1, 2z\right) = \frac{1}{\sqrt{\pi}} e^z (2z)^{-v} K_v(z), \quad (3.42)$$

which gives

$$g_1(4) = \frac{1}{4\sqrt{e}} \left( 5e^{1/4} K_0\left(\frac{1}{4}\right) + e^{1/4} K_1\left(\frac{1}{4}\right) \right).$$

Here, we first show that  $e^z K_0(z)$  and  $e^z K_1(z)$  are decreasing functions with respect to  $z > 0$ . From the definition of  $K_v(z)$  in (3.29), it holds that

$$\frac{d}{dz} K_v(z) = -\frac{1}{2} (K_{v+1}(z) + K_{v-1}(z)),$$

which gives that

$$\begin{aligned} \frac{d}{dz} e^z K_0(z) &= e^z (K_0(z) - K_1(z)), \\ \frac{d}{dz} e^z K_1(z) &= -\frac{1}{2} e^z (K_0(z) - 2K_1(z) + K_2(z)). \end{aligned}$$

From the integral representation of  $K_v(z)$  in (3.29), it holds from  $\cosh 2t = \cosh^2 t - 1$  that

$$\begin{aligned} K_0(z) - K_1(z) &= \int_0^\infty e^{-z \cosh t} (1 - \cosh t) dt < 0 \\ K_0(z) - 2K_1(z) + K_2(z) &= \int_0^\infty e^{-z \cosh t} (\cosh^2 t - \cosh t) dt > 0, \end{aligned}$$

which shows that  $e^z K_0(z)$  and  $e^z K_1(z)$  are decreasing functions with respect to  $z > 0$ .

Then, we obtain

$$\begin{aligned} g_1(4) &= \frac{1}{4\sqrt{e}} \left( 5e^{1/4} K_0\left(\frac{1}{4}\right) + e^{1/4} K_1\left(\frac{1}{4}\right) \right) \\ &\leq \frac{1}{4e^{1/2}} (5e^{0.24} K_0(0.24) + e^{0.24} K_1(0.24)). \end{aligned}$$

By substituting the numerical computation in Fact 3.23, we obtain that

$$\begin{aligned} g_1(4) &\leq \frac{1}{4e^{1/2}} (5e^{0.24} K_0(0.24) + e^{0.24} K_1(0.24)) = 2.27811 \quad \text{to 6S} \\ &< \Gamma(2)B(1/2, 1/2) = \Gamma(2) \frac{\Gamma(1/2)^2}{\Gamma(1)} = \pi, \end{aligned}$$

which concludes the base case of even  $n$  for the reference prior ( $k = 1$ ).

**(1-ii) Induction** Assume that the following holds for some  $m \geq 2$

$$g_1(2m) \leq \Gamma(m)B\left(\frac{1}{2}, \frac{1}{m}\right) = \Gamma(m) \frac{\Gamma(1/2)\Gamma(1/m)}{\Gamma\left(\frac{1}{2} + \frac{1}{m}\right)}.$$

From the definition of  $g_1(\cdot)$  and  $\Gamma(m+1, x) = m\Gamma(m, x) + x^m e^{-x}$ , we have

$$\begin{aligned} g_1(2(m+1)) &= \int_0^1 w^{-\frac{1}{2}} (1-w)^{-1} \Gamma\left(m+1, \frac{1}{2(1-w)}\right) dw \\ &= \int_0^1 w^{-\frac{1}{2}} (1-w)^{-1} \left( m\Gamma\left(m, \frac{1}{2(1-w)}\right) \right. \\ &\quad \left. + (2(1-w))^{-m} e^{-\frac{1}{2(1-w)}} \right) dw \\ &= mg(2m) + \frac{1}{2^m} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-(m+1)} e^{-\frac{1}{2(1-w)}} dw \\ &\leq \Gamma(m+1)B\left(\frac{1}{2}, \frac{1}{m}\right) + \frac{1}{2^m} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-(m+1)} e^{-\frac{1}{2(1-w)}} dw. \end{aligned}$$

Since  $B\left(\frac{1}{2}, \frac{1}{m+1}\right) - B\left(\frac{1}{2}, \frac{1}{m}\right)$  is a decreasing function with respect to  $m > 0$ , we have for  $m \geq 2$ .

$$\begin{aligned} B\left(\frac{1}{2}, \frac{1}{m+1}\right) - B\left(\frac{1}{2}, \frac{1}{m}\right) &= \Gamma\left(\frac{1}{2}\right) \left( \frac{\Gamma\left(\frac{1}{m+1}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{m+1}\right)} - \frac{\Gamma\left(\frac{1}{m}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{m}\right)} \right) \\ &\geq \lim_{s \rightarrow \infty} \Gamma\left(\frac{1}{2}\right) \left( \frac{\Gamma\left(\frac{1}{s+1}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{s+1}\right)} - \frac{\Gamma\left(\frac{1}{s}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{s}\right)} \right) \\ &= \lim_{s \rightarrow \infty} \Gamma\left(\frac{1}{s+1}\right) - \Gamma\left(\frac{1}{s}\right) = 1. \end{aligned}$$

Therefore, it is sufficient to show

$$h(2(m+1)) := \frac{1}{2^m} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-(m+1)} e^{-\frac{1}{2(1-w)}} dw \leq \Gamma(m+1).$$

Again, by letting  $t = \frac{w}{1-w}$ ,  $h$  can be written as

$$\begin{aligned} h(2(m+1)) &= \frac{1}{2^m \sqrt{e}} \int_0^\infty t^{-\frac{1}{2}} (t+1)^{m-\frac{1}{2}} e^{-\frac{t}{t+1}} dt \\ &= \sqrt{\frac{\pi}{e}} \frac{1}{2^m} U\left(\frac{1}{2}, m+1, \frac{1}{2}\right). \end{aligned}$$

From Lemma 3.22, it holds for  $m \geq 2$  that

$$\begin{aligned} h(2(m+1)) &\leq \sqrt{\frac{\pi}{e}} \frac{1}{2^m} \frac{2^{m+1}}{\Gamma\left(\frac{1}{2}\right)} \Gamma\left(m + \frac{1}{2}\right) \\ &= \frac{2}{\sqrt{e}} \Gamma\left(m + \frac{1}{2}\right) \leq \Gamma(m+1), \end{aligned}$$

which concludes the induction when  $n$  is an even number.

**(2) Odd number** Although this case can be easily derived by following the same steps in the case of even numbers, we provide detailed proof for completeness.

**(2-i) Base case  $n = 3$**  From the definition of the upper incomplete gamma function, it holds that

$$\Gamma\left(\frac{3}{2}, x\right) = \frac{\sqrt{\pi}}{2} \operatorname{erfc}(\sqrt{x}) + \sqrt{x} e^{-x},$$

where  $\operatorname{erfc}(\cdot)$  denotes the complementary error function. It is known that the complementary error function is bounded for any  $x \geq 0$  as follows [Simon and Divsalar, 1998]:

$$\operatorname{erfc}(x) \leq e^{-x^2},$$

which gives

$$\Gamma\left(\frac{3}{2}, x\right) \leq \frac{\sqrt{\pi}}{2} e^{-x} + \sqrt{x} e^{-x}.$$



Then, by letting  $t = \frac{w}{1-w}$ , we obtain

$$\begin{aligned}
g_1(3) &\leq \int_0^1 w^{-\frac{1}{2}}(1-w)^{-1} \left( \frac{\sqrt{\pi}}{2} e^{-\frac{1}{2(1-w)}} + \sqrt{\frac{1}{2(1-w)}} e^{-\frac{1}{2(1-w)}} \right) dw \\
&= \int_0^\infty \frac{\sqrt{\pi}}{2\sqrt{e}} (t(t+1))^{-\frac{1}{2}} e^{-\frac{t}{2}} + \frac{1}{\sqrt{2e}} t^{-\frac{1}{2}} e^{-\frac{t}{2}} dt \\
&= \frac{\pi}{2\sqrt{e}} U\left(\frac{1}{2}, 1, \frac{1}{2}\right) + \sqrt{\frac{2}{2e}} \Gamma\left(\frac{1}{2}\right) \\
&= \frac{\pi}{2\sqrt{e}} \frac{e^{1/4}}{\sqrt{\pi}} K_0\left(\frac{1}{4}\right) + \sqrt{\frac{\pi}{e}} \quad \text{by (3.42)} \\
&\leq \frac{1}{2} \sqrt{\frac{\pi}{e}} e^{0.24} K_0(0.24) + \sqrt{\frac{\pi}{e}} = 2.15458 \quad \text{to 6S} \\
&< \frac{\pi}{2} \frac{\Gamma(2/3)}{\Gamma(1.165)} \leq \frac{\pi}{2} \frac{\Gamma(2/3)}{\Gamma(7/6)} = 2.29148 \quad \text{to 6S} \\
&< \Gamma\left(\frac{3}{2}\right) B\left(\frac{1}{2}, \frac{2}{3}\right), \quad (3.43)
\end{aligned}$$

where we substituted the numerical computation in Fact 3.23 and Abramowitz and Stegun [see 1964, 6.1.13 and Tables 6.1] in (3.43) to 6S.

**(2-ii) Induction** Assume that the following holds for some  $m \geq 1$ .

$$g_1(2m+1) \leq \Gamma\left(m + \frac{1}{2}\right) B\left(\frac{1}{2}, \frac{2}{2m+1}\right) = \Gamma\left(m + \frac{1}{2}\right) \frac{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{2}{2m+1}\right)}{\Gamma\left(\frac{1}{2} + \frac{2}{2m+1}\right)}.$$

From the definition and the fact  $\Gamma(s+1, x) = m\Gamma(s, x) + x^s e^{-x}$ , we have

$$\begin{aligned}
g_1(2m+3) &= \int_0^1 w^{-\frac{1}{2}}(1-w)^{-1} \Gamma\left(m + \frac{1}{2} + 1, \frac{1}{2(1-w)}\right) dw \\
&= \int_0^1 w^{-\frac{1}{2}}(1-w)^{-1} \left( m\Gamma\left(m + \frac{1}{2}, \frac{1}{2(1-w)}\right) \right. \\
&\quad \left. + (2(1-w))^{-m-\frac{1}{2}} e^{-\frac{1}{2(1-w)}} \right) dw \\
&= mg(2m+1) + \frac{1}{2^{m+1/2}} \int_0^1 w^{-\frac{1}{2}}(1-w)^{-(m+3/2)} e^{-\frac{1}{2(1-w)}} dw \\
&\leq \Gamma\left(m + \frac{3}{2}\right) B\left(\frac{1}{2}, \frac{2}{2m+1}\right) \\
&\quad + \frac{1}{2^{m+1/2}} \int_0^1 w^{-\frac{1}{2}}(1-w)^{-(m+3/2)} e^{-\frac{1}{2(1-w)}} dw.
\end{aligned}$$

Since

$$\begin{aligned}
B\left(\frac{1}{2}, \frac{2}{2m+3}\right) - B\left(\frac{1}{2}, \frac{2}{2m+1}\right) &= \Gamma\left(\frac{1}{2}\right) \left( \frac{\Gamma\left(\frac{2}{2m+3}\right)}{\Gamma\left(\frac{1}{2} + \frac{2}{2m+1}\right)} - \frac{\Gamma\left(\frac{2}{2m+1}\right)}{\Gamma\left(\frac{1}{2} + \frac{2}{2m}\right)} \right) \\
&\geq 1
\end{aligned}$$

holds for  $m \geq 2$ , it is sufficient to show

$$h(2m+3) := \frac{1}{2^{m+1/2}} \int_0^1 w^{-\frac{1}{2}}(1-w)^{-(m+3/2)} e^{-\frac{1}{2(1-w)}} dw \leq \Gamma\left(m + \frac{3}{2}\right).$$

Again, by letting  $t = \frac{w}{1-w}$ ,  $h(\cdot)$  can be written as

$$\begin{aligned} h(2m+3) &= \frac{1}{2^m \sqrt{e}} \int_0^\infty t^{-\frac{1}{2}} (t+1)^m e^{-\frac{t}{2}} dt \\ &= \sqrt{\frac{\pi}{e}} \frac{1}{2^{m+1/2}} U\left(\frac{1}{2}, m + \frac{3}{2}, \frac{1}{2}\right). \end{aligned}$$

From Lemma 3.22, it holds for all  $m \geq 1$  that

$$\begin{aligned} h(2m+3) &\leq \sqrt{\frac{\pi}{e}} \frac{1}{2^{m+1/2}} \frac{2^{m+3/2}}{\Gamma(\frac{1}{2})} \Gamma(m+1) \\ &= \frac{2}{\sqrt{e}} \Gamma(m+1) \leq \Gamma\left(m + \frac{3}{2}\right). \end{aligned}$$

The proof of Lemma 3.21 for the case of  $k = 1$  is complete.

**For the Jeffreys prior** ( $k = 2$ )

The proofs here shares the same steps to that for the reference prior.

**(1) Even number** Since  $n_0 = (2, 4 - k)$ , we have to consider  $n = 2$  as a base case.

**(1-i) Base case**  $n = 2$  From the definition of the upper incomplete gamma function, it holds that

$$\Gamma(1, x) = e^{-x}.$$

By letting  $t = \frac{w}{1-w}$ , we have

$$\begin{aligned} g_2(2) &= \frac{1}{\sqrt{e}} \int_0^\infty e^{-\frac{t}{2}} t^{-\frac{1}{2}} dt = \sqrt{\frac{2}{e}} \Gamma\left(\frac{1}{2}\right) \\ &\leq \sqrt{\pi} e^{-1/6} 2^{1/4} \\ &\leq \Gamma\left(\frac{1}{2}\right) \frac{\Gamma\left(\frac{1}{\sqrt{2}}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{\sqrt{2}}\right)} = \Gamma(1) B\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right), \end{aligned}$$

where we applied Lemma 3.18 in the last inequality.

**(1-ii) Induction** Assume that the following holds for some  $m \geq 1$

$$g_2(2m) \leq \Gamma(m) B\left(\frac{1}{2}, \frac{1}{m\sqrt{2m}}\right).$$

From the definition and the fact  $\Gamma(m+1, x) = m\Gamma(m, x) + x^m e^{-x}$ , we have

$$\begin{aligned} g_2(2(m+1)) &= \int_0^1 w^{-\frac{1}{2}} (1-w)^{-\frac{3}{2}} \Gamma\left(m+1, \frac{1}{2(1-w)}\right) dw \\ &= \int_0^1 w^{-\frac{1}{2}} (1-w)^{-\frac{3}{2}} \left( m\Gamma\left(m, \frac{1}{2(1-w)}\right) \right. \\ &\quad \left. + (2(1-w))^{-m} e^{-\frac{1}{2(1-w)}} \right) dw \\ &= mg_2(2m) + \frac{1}{2^m} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-(m+\frac{3}{2})} e^{-\frac{1}{2(1-w)}} dw \\ &\leq \Gamma(m+1) B\left(\frac{1}{2}, \frac{1}{m\sqrt{2m}}\right) + \frac{1}{2^m} \int_0^1 w^{-\frac{1}{2}} (1-w)^{-(m+\frac{3}{2})} e^{-\frac{1}{2(1-w)}} dw. \end{aligned}$$

Here, it holds for  $m \geq 1$  that

$$B\left(\frac{1}{2}, \frac{1}{(m+1)\sqrt{2(m+1)}}\right) - B\left(\frac{1}{2}, \frac{1}{m\sqrt{2m}}\right) \geq \sqrt{2m+2}.$$

Therefore, it is sufficient to show

$$h(2(m+1)) := \frac{1}{2^m} \int_0^1 w^{-\frac{1}{2}}(1-w)^{-(m+\frac{3}{2})} e^{-\frac{1}{2(1-w)}} dw \leq \sqrt{2(m+1)}\Gamma(m+1).$$

Again, by letting  $t = \frac{w}{1-w}$ ,  $h$  can be written as

$$\begin{aligned} h(2(m+1)) &= \frac{1}{2^m \sqrt{e}} \int_0^\infty t^{-\frac{1}{2}}(t+1)^m e^{-\frac{t}{2}} dt \\ &= \sqrt{\frac{\pi}{e}} \frac{1}{2^m} U\left(\frac{1}{2}, m + \frac{3}{2}, \frac{1}{2}\right). \end{aligned}$$

From Lemma 3.22, it holds for  $m \geq 1$  that

$$\begin{aligned} h(2(m+1)) &\leq \sqrt{\frac{\pi}{e}} \frac{1}{2^m} \frac{2^{m+\frac{3}{2}}}{\Gamma(\frac{1}{2})} \Gamma(m+1) \\ &= \frac{2\sqrt{2}}{\sqrt{e}} \Gamma(m+1) \leq \sqrt{2(m+1)}\Gamma(m+1), \end{aligned}$$

which concludes the induction when  $n$  is an even number.

**(2) Odd number** Although this case can be easily derived by following the same steps in the case of even numbers, we provide detailed proof for completeness.

**(2-i) Base case  $n = 3$**  From the definition of the upper incomplete gamma function, it holds that

$$\Gamma\left(\frac{3}{2}, x\right) = \frac{\sqrt{\pi}}{2} \operatorname{erfc}(\sqrt{x}) + \sqrt{x}e^{-x},$$

where  $\operatorname{erfc}(\cdot)$  denotes the complementary error function. It is known that the complementary error function is bounded for any  $x \geq 0$  as follows [Simon and Divsalar, 1998]:

$$\operatorname{erfc}(x) \leq e^{-x^2},$$

which gives

$$\Gamma\left(\frac{3}{2}, x\right) \leq \frac{\sqrt{\pi}}{2} e^{-x} + \sqrt{x}e^{-x}.$$

Then, by letting  $t = \frac{w}{1-w}$ , we obtain

$$\begin{aligned} g_2(3) &\leq \int_0^1 w^{-\frac{1}{2}}(1-w)^{-\frac{3}{2}} \left( \frac{\sqrt{\pi}}{2} e^{-\frac{1}{2(1-w)}} + \sqrt{\frac{1}{2(1-w)}} e^{-\frac{1}{2(1-w)}} \right) dw \\ &= \int_0^\infty \frac{\sqrt{\pi}}{2\sqrt{e}} t^{-\frac{1}{2}} e^{-\frac{t}{2}} + \frac{1}{\sqrt{2e}} t^{-\frac{1}{2}}(1+t)^{\frac{1}{2}} e^{-\frac{t}{2}} dt \\ &= \sqrt{\frac{\pi}{2e}} \Gamma\left(\frac{1}{2}\right) + \sqrt{\frac{\pi}{2e}} U\left(\frac{1}{2}, 2, \frac{1}{2}\right) \\ &= \frac{\pi}{\sqrt{2e}} + \frac{1}{2\sqrt{2e}} \left( e^{1/4} K_0\left(\frac{1}{4}\right) + e^{1/4} K_1\left(\frac{1}{4}\right) \right) \quad \text{by (3.41) and (3.42)} \\ &\leq \frac{\pi}{\sqrt{2e}} + \frac{1}{2\sqrt{2e}} (e^{0.24} K_0(0.24) + e^{0.24} K_1(0.24)) = 2.84642 \quad \text{to 6S} \\ &< \Gamma\left(\frac{3}{2}\right) B\left(\frac{1}{2}, \frac{2}{3\sqrt{3}}\right) = 3.35278 \end{aligned}$$

where we substituted the numerical computation in Fact 3.23.

**(2-ii) Induction** Assume that the following holds for some  $m \geq 1$ .

$$g_2(2m+1) \leq \Gamma\left(m + \frac{1}{2}\right) B\left(\frac{1}{2}, \frac{2}{(2m+1)\sqrt{2m+1}}\right).$$

From the definition and the fact  $\Gamma(s+1, x) = m\Gamma(s, x) + x^s e^{-x}$ , we have

$$\begin{aligned} g_2(2m+3) &= \int_0^1 w^{-\frac{1}{2}}(1-w)^{-\frac{3}{2}} \Gamma\left(m + \frac{1}{2} + 1, \frac{1}{2(1-w)}\right) dw \\ &= \int_0^1 w^{-\frac{1}{2}}(1-w)^{-\frac{3}{2}} \left( m\Gamma\left(m + \frac{1}{2}, \frac{1}{2(1-w)}\right) \right. \\ &\quad \left. + (2(1-w))^{-m-\frac{1}{2}} e^{-\frac{1}{2(1-w)}} \right) dw \\ &= mg_2(2m+1) + \frac{1}{2^{m+1/2}} \int_0^1 w^{-\frac{1}{2}}(1-w)^{-(m+2)} e^{-\frac{1}{2(1-w)}} dw \\ &\leq \Gamma\left(m + \frac{3}{2}\right) B\left(\frac{1}{2}, \frac{2}{(2m+1)\sqrt{2m+1}}\right) \\ &\quad + \frac{1}{2^{m+1/2}} \int_0^1 w^{-\frac{1}{2}}(1-w)^{-(m+2)} e^{-\frac{1}{2(1-w)}} dw. \end{aligned}$$

Since

$$B\left(\frac{1}{2}, \frac{2}{(2m+3)\sqrt{2m+3}}\right) - B\left(\frac{1}{2}, \frac{2}{(2m+1)\sqrt{2m+1}}\right) \geq \sqrt{2m+3}$$

holds for  $m \geq 1$ , it is sufficient to show

$$h(2m+3) := \frac{1}{2^{m+1/2}} \int_0^1 w^{-\frac{1}{2}}(1-w)^{-(m+3/2)} e^{-\frac{1}{2(1-w)}} dw \leq \sqrt{2m+3} \Gamma\left(m + \frac{3}{2}\right).$$

Again, by letting  $t = \frac{w}{1-w}$ ,  $h(\cdot)$  can be written as

$$\begin{aligned} h(2m+3) &= \frac{1}{2^{m+1/2}\sqrt{e}} \int_0^\infty t^{-\frac{1}{2}}(t+1)^{m+\frac{1}{2}} e^{-\frac{t}{2}} dt \\ &= \sqrt{\frac{\pi}{e}} \frac{1}{2^{m+1/2}} U\left(\frac{1}{2}, m+2, \frac{1}{2}\right). \end{aligned}$$

From Lemma 3.22, it holds for all  $m \geq 1$  that

$$\begin{aligned} h(2m+3) &\leq \sqrt{\frac{\pi}{e}} \frac{1}{2^{m+1/2}} \frac{2^{m+2}}{\Gamma\left(\frac{1}{2}\right)} \Gamma\left(m + \frac{3}{2}\right) \\ &= \frac{2\sqrt{2}}{\sqrt{e}} \Gamma\left(m + \frac{3}{2}\right) \leq \sqrt{2m+3} \Gamma\left(m + \frac{3}{2}\right). \end{aligned}$$

The proof of Lemma 3.21 for the case of  $k = 2$  is complete.  $\square$

*Proof of Lemma 3.22.* Similarly to the proof of Lemma 3.21, we apply mathematical induction.

**Base case:**  $b = 2$  When  $b = 2$  ( $m = 4$ ), it holds from (3.41) and (3.42) that

$$\begin{aligned}
U\left(\frac{1}{2}, 2, \frac{1}{2}\right) &= \frac{1}{4}U\left(\frac{3}{2}, 2, \frac{1}{2}\right) + \frac{1}{2}U\left(\frac{1}{2}, 1, \frac{1}{2}\right) \\
&= \frac{e^{1/4}}{2\sqrt{\pi}} \left( K_0\left(\frac{1}{4}\right) + K_1\left(\frac{1}{4}\right) \right) \\
&\leq \frac{1}{2\sqrt{\pi}} (e^{0.24} K_0(0.24) + e^{0.24} K_1(0.24)) = 1.97198 \quad \text{to 6S} \\
&< \frac{4}{\Gamma(1/2)} \Gamma\left(\frac{3}{2}\right) = 2,
\end{aligned}$$

where we substituted the numerical computation to 6S given in Fact 3.23. When  $b = 2 + \frac{1}{2}$  ( $m = 5$ ), it holds that

$$U\left(\frac{1}{2}, 2 + \frac{1}{2}, \frac{1}{2}\right) = 2\sqrt{2} < \frac{4\sqrt{2}}{\Gamma(1/2)} \Gamma(2) = 4\sqrt{\frac{2}{\pi}}.$$

**Induction** For the confluent hypergeometric function of the second kind, the following recurrence relation holds as follows [Olver et al., 2010, 13.3.8]

$$(b - a - 1)U(a, b - 1, z) + (1 - b - z)U(a, b, z) + zU(a, b + 1, z) = 0.$$

Injecting  $a, z = \frac{1}{2}$  gives

$$\begin{aligned}
U\left(\frac{1}{2}, b + 1, \frac{1}{2}\right) &= (2b - 1)U\left(\frac{1}{2}, b, \frac{1}{2}\right) - (2b - 3)U\left(\frac{1}{2}, b - 1, \frac{1}{2}\right) \\
&\leq (2b - 1)U\left(\frac{1}{2}, b, \frac{1}{2}\right).
\end{aligned}$$

Therefore, if

$$U\left(\frac{1}{2}, b, \frac{1}{2}\right) \leq \frac{2^b}{\Gamma\left(\frac{1}{2}\right)} \Gamma\left(b - \frac{1}{2}\right)$$

holds, then we obtain

$$\begin{aligned}
U\left(\frac{1}{2}, b + 1, \frac{1}{2}\right) &\leq (2b - 1) \frac{2^b}{\Gamma\left(\frac{1}{2}\right)} \Gamma\left(b - \frac{1}{2}\right) \\
&= \frac{2^{b+1}}{\Gamma\left(\frac{1}{2}\right)} \Gamma\left(b + \frac{1}{2}\right).
\end{aligned}$$

The proof of Lemma 3.22 is complete.  $\square$

### 3.7.9 Proof of the suboptimality of TS

In this section, we provide proof of the suboptimality of TS with  $k \geq 1$  for the uniform bandits with unknown supports.

*Proof of Theorem 3.2.* Since TS-T starts from playing every arms twice,  $N_i(s) \geq 2$  holds for all  $i \in \{1, 2\}$  and  $s \geq 5$ . Then, it holds for  $T \geq 5$  that

$$\begin{aligned}
\mathbb{E}[\text{Reg}(T)] &= \Delta_2 \mathbb{E} \left[ \sum_{t=1}^T \mathbb{1}[i(t) = 2] \right] \\
&\geq \Delta_2 \mathbb{E} \left[ \sum_{t=5}^T \mathbb{1}[i(t) = 2, N_1(t) = 2] \right].
\end{aligned}$$

Since  $N_1(t)$  denotes the number of playing arm 1 until round  $t$ , if an event  $\{i(s) \neq 2, N_1(s) = 2\}$  occurs for some  $s \geq 5$ , then  $N_1(t) > 2$  holds for  $t > s$ . Therefore, for any  $t \geq 5$ ,

$$\begin{aligned} \{i(t) = 2, N_1(t) = 2\} &\Leftrightarrow \{\forall s \in [1, t-4] : i(s+4) = 2\} \\ &\Leftrightarrow \{\forall s \in [1, t-4] : \tilde{\mu}_1(s+4) < \mu_2\}. \end{aligned}$$

By letting  $T' = T - 4$ , we have

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=5}^T \mathbb{1}[i(t) = 2, N_1(t) = 2] \right] &= \mathbb{E} \left[ \sum_{t=5}^T \mathbb{1}[\forall s \in [1, t-4] : \tilde{\mu}_1(s+4) < \mu_2] \right] \\ &= \mathbb{E}_{x^{(1)}, x^{(2)}} \left[ \sum_{s=1}^{T'} \left( \mathbb{P}[\tilde{\mu}_1 \leq \mu_2 \mid x_1^{(1)}, x_1^{(2)}] \right)^s \right]. \end{aligned}$$

Since  $\tilde{\mu}_1 | \tilde{\sigma}_1 \sim \text{Uniform}_{\mu\sigma}(\hat{\mu}_{1,2}, \tilde{\sigma}_1 - \hat{\sigma}_{1,2})$ , if  $\hat{\mu}_{1,2} + \frac{\tilde{\sigma}_1 - \hat{\sigma}_{1,2}}{2} \leq \mu_2$  holds, then  $\tilde{\mu}_1 \leq \mu_2$  always holds since  $\tilde{\mu}_1$  is generated from the fixed posterior distribution. Therefore, we have

$$\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 \mid x_1^{(1)}, x_1^{(2)}] \geq \mathbb{1}[\hat{\mu}_{1,2} \leq \mu_2] \mathbb{P}[\tilde{\sigma}_1 \leq 2(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2} \mid x_1^{(1)}, x_1^{(2)}],$$

since  $\tilde{\sigma}_1 \geq \hat{\sigma}_{1,2}$  holds. Therefore, we obtain that

$$\begin{aligned} &\mathbb{P}[\tilde{\sigma}_1 \leq 2(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2} \mid x_1^{(1)}, x_1^{(2)}] \\ &= \mathbb{1}[\hat{\mu}_{1,2} \leq \mu_2] \int_{\hat{\sigma}_{1,2}}^{(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2}} k(k+1) \hat{\sigma}_{1,2}^k \frac{s - \hat{\sigma}_{1,2}}{s^{k+2}} ds \\ &= \mathbb{1}[\hat{\mu}_{1,2} \leq \mu_2] \left( 1 - (k+1) \left( \frac{\hat{\sigma}_{1,2}}{2(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2}} \right)^k \right. \\ &\quad \left. + k \left( \frac{\hat{\sigma}_{1,2}}{2(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2}} \right)^{k+1} \right) \\ &\geq \mathbb{1}[\hat{\mu}_{1,2} \leq \mu_2] \left( 1 - (k+1) \left( \frac{\hat{\sigma}_{1,2}}{2(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2}} \right)^k \right) \\ &\geq \mathbb{1}[x_1^{(2)} \leq \mu_2] \left( 1 - (k+1) \left( \frac{\hat{\sigma}_{1,2}}{2(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2}} \right)^k \right) \quad \text{by } x_1^{(2)} \geq \hat{\mu}_{1,2} \end{aligned}$$

For simplicity, let us define

$$\begin{aligned} q_n(k) &= q(k | x_1^{(1)}, x_1^{(2)}) = \mathbb{1}[x_1^{(2)} \leq \mu_2] \left( 1 - (k+1) \left( \frac{\hat{\sigma}_{1,2}}{2(\mu_2 - \hat{\mu}_{1,2}) + \hat{\sigma}_{1,2}} \right)^k \right) \\ &= \mathbb{1}[x_1^{(2)} \leq \mu_2] \left( 1 - (k+1) \left( \frac{x_1^{(2)} - x_1^{(1)}}{2(\mu_2 - x_1^{(1)})} \right)^k \right). \end{aligned}$$

Then, it holds that

$$\begin{aligned}
& \mathbb{E}_{x^{(1)}, x^{(2)}} \left[ \sum_{s=1}^{T'} \left( \mathbb{P} \left[ \tilde{\mu}_1 \leq \mu_2 \mid x_1^{(1)}, x_1^{(2)} \right] \right)^s \right] \\
& \geq \mathbb{E}_{x^{(1)}, x^{(2)}} \left[ \sum_{s=1}^{T'} (q_n(k))^s \right] \\
& = \mathbb{E}_{x^{(1)}, x^{(2)}} \left[ \left( 1 - (q_n(k))^{T'} \right) \frac{q_n(k)}{1 - q_n(k)} \right] \\
& \geq \frac{1}{2} \mathbb{E}_{x^{(1)}, x^{(2)}} \left[ \mathbb{1} [x_1^{(2)} \leq \mu_2, (q_n(k))^{T'} \leq 1/2] \frac{q_n(k)}{1 - q_n(k)} \right] \\
& \geq \frac{1}{2} \mathbb{E}_{x^{(1)}, x^{(2)}} \left[ \frac{\mathbb{1} [x_1^{(2)} \leq \mu_2, (q_n(k))^{T'} \leq 1/2]}{(k+1) \left( \frac{x_1^{(2)} - x_1^{(1)}}{2(\mu_2 - x_1^{(1)})} \right)^k} \right] - \frac{1}{2}.
\end{aligned}$$

Here, it holds that

$$\begin{aligned}
(q_n(k))^{T'} \leq 1/2 & \Leftrightarrow \left( 1 - (k+1) \left( \frac{x_1^{(2)} - x_1^{(1)}}{2(\mu_2 - x_1^{(1)})} \right)^k \right)^{T'} \leq \frac{1}{2} \\
& \Leftrightarrow 1 - (k+1) \left( \frac{x_1^{(2)} - x_1^{(1)}}{2(\mu_2 - x_1^{(1)})} \right)^k \leq 2^{-\frac{1}{T'}} \\
& \Leftrightarrow 1 - 2^{-\frac{1}{T'}} \leq (k+1) \left( \frac{x_1^{(2)} - x_1^{(1)}}{2(\mu_2 - x_1^{(1)})} \right)^k \\
& \Leftrightarrow \frac{\log 2}{T'} \leq (k+1) \left( \frac{x_1^{(2)} - x_1^{(1)}}{2(\mu_2 - x_1^{(1)})} \right)^k.
\end{aligned}$$

From Lemma 3.12 with  $n = 2$ , it holds that

$$\begin{aligned}
& \mathbb{E}_{x^{(1)}, x^{(2)}} \left[ \frac{\mathbb{1} [x_1^{(2)} \leq \mu_2, (q_n(k))^{T'} \leq 1/2]}{(k+1) \left( \frac{x_1^{(2)} - x_1^{(1)}}{2(\mu_2 - x_1^{(1)})} \right)^k} \right] \\
& = \iint_{\substack{a_1 \leq y \leq z \leq \mu_2, \\ \frac{\log 2}{(k+1)T'} \leq \left( \frac{z-y}{2(\mu_2 - y)} \right)^k}} 2\sigma_2^{-2} \frac{(2(\mu_2 - y))^k}{(k+1)(z-y)^k} dz dy \\
& = \frac{2}{\sigma_2^2} \int_{a_1}^{\mu_2} \int_{y+2(\mu_2-y)B_k}^{\mu_2} \frac{(2(\mu_2 - y))^k}{(k+1)(z-y)^k} dz dy,
\end{aligned}$$

where we denoted  $B_k = \left( \frac{\log 2}{T'(k+1)} \right)^{1/k}$ . Then, by direct computation, we obtain for  $k = 1$

$$\begin{aligned}
\frac{2}{\sigma_2^2} \int_{a_1}^{\mu_2} \int_{y+2(\mu_2-y)A_1}^{\mu_2} \frac{2(\mu_2 - y)}{2(z-y)} dz dy & = \frac{1}{\sigma_2^2} \int_{a_1}^{\mu_2} 2(\mu_2 - y) \log \left( \frac{2T'}{\log 2} \right) dy \\
& = \frac{1}{\sigma_2^2} (\mu_2 - a_1)^2 \log \left( \frac{2T'}{\log 2} \right) \\
& = \frac{1}{4} \log \left( \frac{2T'}{\log 2} \right) \tag{3.44}
\end{aligned}$$

and for  $k \geq 2$  that

$$\begin{aligned}
& \frac{2}{\sigma_2^2} \int_{a_1}^{\mu_2} \int_{y+2(\mu_2-y)B_k}^{\mu_2} \frac{(2(\mu_2-y))^k}{(k+1)(z-y)^k} dz dy \\
&= \frac{2}{(k^2-1)\sigma_2^2} \int_{a_1}^{\mu_2} 2(\mu_2-y) \left( \frac{1}{A^{k-1}} - 2^{k-1} \right) dy \\
&= \frac{2}{(k^2-1)\sigma_2^2} (\mu_2 - a_1)^2 \left( \frac{1}{B_k^{k-1}} - 2^{k-1} \right) \\
&= \frac{1}{2(k^2-1)} \left( \left( \frac{(k+1)T'}{\log 2} \right)^{\frac{k-1}{k}} - 2^{k-1} \right), \quad (3.45)
\end{aligned}$$

where (3.44) and (3.45) hold from the assumption

$$a_1 = a_2, \quad \text{and} \quad (\mu_2, \sigma_2) = \left( \frac{a_2 + b_2}{2}, b_2 - a_2 \right).$$

Note that for  $T' \geq 1$  and  $k \geq 2$ ,  $B_k < \frac{1}{2}$  holds.  $\square$

### 3.8 Conclusion

In this chapter, we demonstrated the importance of choosing noninformative priors for the vanilla TS under the uniform bandit models with unknown supports. Although the uniform prior was found to be the optimal choice in terms of the expected problem-dependent regret, we showed that the use of the uniform prior is problematic since it does not maintain invariance under reparameterizations. In particular, we observed that TS with the uniform prior leads to varying results for different parameterizations, which makes the optimality under the specific parameterization less informative in general. Therefore, if an agent considers different parameterizations for the uniform models under TS with the uniform prior, they need to check its optimality and may need to modify their data processing accordingly.

To address such limitations, we proposed a variant of TS, Thompson sampling with truncation (TS-T), where the parameters are sampled from a pre-processed posterior distribution. We showed that TS-T with the invariant priors could achieve the optimal regret bound asymptotically for both the uniform models and the Gaussian models, whereas the same choice of the invariant priors was suboptimal under the vanilla TS. Our analysis was supported by the simulation results where the invariant priors under TS-T showed a better performance than those under the vanilla TS.

In summary, this chapter demonstrated the importance of noninformative priors that are invariant under one-to-one reparameterization of parameters in the context of the MAB problems for the multiparameter models. One interesting observation is the similarity of the behavior of TS for the uniform bandits and the Gaussian bandits. Since both the uniform distributions and the Gaussian distributions belong to the univariate LS family, we expect that the analysis of TS can be extended to all distributions in the location-scale (LS) family.



## Chapter 4

### Pareto Bandits

In the previous chapter, we considered bandit models where the reward distributions belong to the univariate location-scale (LS) family. Although we extended the understanding of Thompson Sampling (TS) to uniform bandits, the optimality of TS has only been examined for distributions possessing a light tail, where the moment-generating function exists. In this chapter, we study the optimality of TS for the Pareto model that has a heavy tail and is a non-regular model. We first derive the closed form of the problem-dependent constant that appears in the theoretical lower bound in Pareto models, which is not trivial, unlike those for exponential families. Based on this result, the study establishes that certain probability matching priors enable TS to achieve optimal regret bounds, which is the first result for two-parameter Pareto bandit models, to this author's knowledge. In contrast, we find that TS with other probability matching priors, such as the Jeffreys priors and the reference priors, exhibit a similar suboptimality as observed in previous two-parameter bandit models. However, this study also demonstrates that TS with the reference prior and the Jeffrey prior can achieve the asymptotic lower bound if one utilizes a truncation procedure proposed in the previous chapter. The findings of this study highlight the importance of selecting noninformative priors and demonstrate the effectiveness of truncation procedures not only in light-tailed but also in heavy-tailed bandit models.

#### 4.1 Introduction

In this section, we provide a brief introduction to the background and motivation behind this chapter. We also summarize the contributions made in this chapter.

##### 4.1.1 Chapter background

In multi-armed bandit (MAB) problems, a large number of studies have considered the problems under light-tailed distributions where the moment-generating function exists. For example, one can find the studies targeting the Bernoulli distribution [Kaufmann et al., 2012b], the Gaussian distribution [Auer et al., 2002, Honda and Takemura, 2014], the Laplace distribution [Lai and Robbins, 1985], and distributions in the one-dimensional exponential family [Korda et al., 2013].

In practice, however, heavy-tailed distributions that do not have a finite exponential moment have been widely adopted for the analysis of many stochastic systems. Specifically, there is statistical or experimental evidence that supports the appropriateness of heavy-tailed distributions in quantum physics [Khalfin, 1958, Wilkinson et al., 1997], natural phenomenon description [Córdoba, 2008, Van Montfort and Witter, 1986], and economics [Bradley and Taqqu, 2003, Mandelbrot, 1960, Mittnik et al., 1998, Oancea et al., 2017]. Notable examples of heavy-tailed distributions are Pareto distributions (and

other power-law distributions), lognormal distributions, Weibull distributions (with the shape parameter less than 1), and non-Gaussian stable distributions. Although stable distributions are a rich class of distributions that can be formulated by several mathematical properties such as non-zero skewness and heavy-tailedness, only the Gaussian, Cauchy, and Lévy distributions are known to have closed forms. Other stable distributions can only be expressed in terms of the characteristic function [Nolan, 2020]. Notice that most distributions considered in practice belong to one of the aforementioned distributions [Foss et al., 2011, Nolan, 2020].

In MAB literature, Bubeck et al. [2013] showed that MAB under heavy-tailed distributions could achieve the same optimal regret as that under sub-Gaussian distributions if distributions admit the finite variance. Based on this result, several studies have considered the MAB problem under the heavy-tailed distributions with the infinite variance [Medina and Yang, 2016, Shao et al., 2018, Xue et al., 2020]. Although algorithms with assumptions on the moments can deal with several heavy-tailed distributions, Pareto distributions, and Student’s  $t$ -distributions are the main target distributions in experiments [Agrawal et al., 2021b, Lee et al., 2020]. This is because well-utilized distributions such as Weibull, lognormal, and Gumbel distributions always have finite variance, for which the optimality is already known.

In this chapter, we investigate MAB problems under the type-1 Pareto distribution that has a power-law property without any other assumptions on moments except the finite first raw moment assumption, i.e., finite mean. The finite mean assumption is necessary to define regret, which implies that this assumption is unavoidable to define the bandit problem. Note that the power-law properties are ubiquitous in many applications with high-variance distributions, such as social sciences [Mahanti et al., 2013], economics [Nirei and Aoki, 2016, Oancea et al., 2017], and size distributions [Barro and Jin, 2011, Córdoba, 2008]. Therefore, MAB problems under Pareto distributions or further power-law distributions are worth considering in practice. Since Pareto distributions admit moments of order smaller than its shape parameter, the assumption on the moments implies an assumption on its shape parameter, which is not usually given in practice [Clauset et al., 2007]. Notice that the raw moments of the Pareto distribution are expressed by scale and shape parameters. Thus, the assumption of a uniform bound on the raw moments implies an assumption on both parameters of Pareto distributions, which is more restrictive.

It is worth mentioning that when the scale parameter is known, then the Pareto distribution falls into the one-parameter exponential family, where TS with Jeffreys prior shows asymptotic optimality when the variance is finite [Korda et al., 2013]. However, when the scale parameter is unknown, the Pareto distributions are not in the exponential family and even do not satisfy the Fisher information (FI) regularity condition. Since the support of Pareto distributions depends on the scale parameter, inaccurate assumptions on the scale parameter would have a large impact on the performance of the algorithms.

#### 4.1.2 Chapter contribution

This chapter presents new findings on the performance of TS in bandit models with heavy-tailed and multiparameter reward distributions, specifically in the two-parameter Pareto bandits. We demonstrate that TS with certain probability matching priors can achieve optimal regret bounds, extending the understanding of TS beyond light-tailed distributions or exponential families. This is the first result not only for bandit algorithms in the two-parameter Pareto bandit models but also for TS-based algorithms in the multiparameter heavy-tailed bandit models, to this author’s knowledge. We further show that TS with different choices of probability matching priors, including the reference priors and the Jeffreys prior, cannot achieve the optimal regret bounds. This result

is consistent with previous findings on the optimality of TS in multiparameter bandits. In addition, we show that the instance of TS with truncation (TS-T) proposed in Chapter 3 can achieve the optimal regret bound with the Jeffreys prior or the reference prior. This result suggests that one can design optimal TS-based algorithms with the reference prior and a truncation procedure rather than relying on the search for suitable priors in other bandit models.

The contributions of this chapter are summarized as follows:

- We prove the asymptotic optimality/suboptimality of TS with noninformative priors under probability matching criteria, emphasizing the importance of selecting appropriate priors in the context of two-parameter Pareto models.
- We demonstrate the effectiveness of a truncation procedure to the parameter space of the posterior distribution in the heavy-tailed bandits. Specifically, we prove the optimality of TS-T in the Pareto bandits with the reference prior and the Jeffreys prior.
- We solve the MAB problem under a two-parameter Pareto distribution without any additional assumptions on moments and parameters, which provides a new perspective other than the moment angle to study the heavy-tailed MABs.

### 4.1.3 Chapter organization

The organization of the rest of this chapter is as follows. We formulate the stochastic MAB problems under the Pareto distribution and derive its regret lower bound in Section 4.2. We also briefly explain the bounded moment bandits in the same section. Then, in Section 4.3, we formulate TS for the Pareto bandits based on the probability matching priors and their corresponding posteriors. To address the suboptimality issue of the Jeffreys prior and the reference prior, we extend the TS-T strategy presented in Section 3.3.2 to the Pareto models. In Section 4.4, we provide the main results on the optimality of TS and TS-T, whose detailed proofs are given in Section 4.6. Numerical results that support our theoretical analysis are provided in Section 4.5.

## 4.2 Problem Formulation

In this section, we formulate the stochastic  $K$ -armed Pareto bandit problems, in which the reward associated with each arm is generated from the corresponding Pareto distribution with fixed parameters. We derive the exact form of the problem-dependent constant that appears in the lower bound of the expected regret in Pareto bandits.

### 4.2.1 Pareto distribution

In  $K$ -armed Pareto bandit problem, an agent chooses an arm  $i \in [K]$  at each round  $t \in \mathbb{N}$  and observes an independent and identically distributed reward from  $\text{Pareto}(\sigma_i, \alpha_i)$ , where  $\text{Pareto}(\sigma, \alpha)$  denotes the Pareto distribution parameterized by scale  $\sigma > 0$  and shape  $\alpha > 0$ . This has the density function of form

$$f_{\sigma, \alpha}^{\text{Pa}}(x) = \frac{\alpha \sigma^\alpha}{x^{\alpha+1}} \mathbb{1}[x \geq \sigma], \quad (4.1)$$

where  $\mathbb{1}[\cdot]$  denotes the indicator function. Since the support of the density depends on the scale parameter, two-parameter Pareto distributions do not belong to the well-studied distributions such as the exponential family, unlike one-parameter Pareto distributions where the scale parameter is fixed.

For a random variable  $X$ , let  $\gamma^* = \sup\{\gamma > 0 : \nu_\gamma < \infty\}$  denote the supremum order of central moments  $\nu_\gamma = \mathbb{E}[|X|^\gamma]$  that a distribution of  $X$  admits. Then, the reward distribution of an arm will admit  $\gamma$ -th moments for any  $\gamma < \gamma^*$ . Note that  $\gamma^* \leq 1$  and  $\gamma^* \leq 2$  implies that a distribution has infinite mean and infinite variance, respectively. As  $\text{Pareto}(\sigma, \alpha)$  admits  $\gamma$ -th moments,  $\nu_\gamma = \frac{\sigma^\gamma \alpha}{\alpha - \gamma}$  for all  $\gamma < \alpha$  and have infinite moments for  $\gamma \geq \alpha$ ,  $\gamma^* = \alpha$  holds for the Pareto distribution. Therefore,  $\alpha > 1$  is a necessary condition of an arm to have a finite mean  $\mu(\theta) = \nu_1 = \frac{\sigma \alpha}{\alpha - 1}$  for  $\theta = (\sigma, \alpha)$ , which is required to define the sub-optimality gap  $\Delta_i := \max_{j \in [K]} \mu_j - \mu_i$ . We assume without loss of generality that arm 1 has the maximum mean for simplicity, i.e.,  $\mu_1 = \max_{i \in [K]} \mu_i$  and  $\Delta_i = \mu_1 - \mu_i$ . In this chapter,  $\nu_{i,\gamma}$  denotes the  $\gamma$ -th moment of the arm  $i \in [K]$ .

#### 4.2.2 Asymptotic regret lower bound

The problem-dependent regret lower bound in (2.1) can be written as

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \geq \sum_{i=2}^K \frac{\Delta_i}{\inf_{(\sigma, \alpha): \mu(\theta) > \mu_1} \text{KL}(\text{Pareto}(\sigma_i, \alpha_i), \text{Pareto}(\sigma, \alpha))}. \quad (4.2)$$

The KL divergence between Pareto distributions is given as

$$\begin{aligned} & \text{KL}(\text{Pareto}(\sigma_1, \alpha_1), \text{Pareto}(\sigma_2, \alpha_2)) \\ &= \begin{cases} \log\left(\frac{\alpha_1}{\alpha_2}\right) + \alpha_2 \log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{\alpha_2}{\alpha_1} - 1 & \text{if } \sigma_2 \leq \sigma_1, \\ \infty & \text{otherwise.} \end{cases} \end{aligned}$$

Here the divergence sometimes becomes infinite since the scale parameter  $\sigma$  determines the support of the Pareto distribution. We denote the numerator in (4.2) for  $i \neq 1$  by

$$\begin{aligned} \text{KL}_{\inf}(i) &:= \inf_{\theta: \mu(\theta) > \mu_1} \text{KL}(\text{Pareto}(\sigma_i, \alpha_i), \text{Pareto}(\theta)) \\ &= \inf_{\theta \in \Theta_i} \log \frac{\alpha_i}{\alpha} + \alpha \log \frac{\sigma_i}{\sigma} + \frac{\alpha}{\alpha_i} - 1, \end{aligned}$$

where

$$\Theta_i = \{(\sigma, \alpha) \in (0, \sigma_i] \times (0, \infty) : \mu(\sigma, \alpha) > \mu_1\}. \quad (4.3)$$

Notice that  $\Theta_i$  allows parameters whose expected rewards are infinite ( $\alpha \in (0, 1]$ ), although we consider a bandit model with  $\alpha_i > 1$  for all  $i \in [K]$  so that the sub-optimality gap  $\Delta_i$  becomes finite. This implies that  $\text{KL}_{\inf}(i)$  does not depend on whether the agent considers the possibility that an arm has the infinite expected reward or not. Then, we can simply rewrite the lower bound in (4.2) as

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \geq \sum_{i=2}^K \frac{\Delta_i}{\text{KL}_{\inf}(i)}.$$

The following lemma shows the closed form of this infimum, whose proof is given in Section 4.6.1.

**Lemma 4.1.** *For any arm  $i \neq 1$ , it holds that*

$$\text{KL}_{\inf}(i) = \log \left( \alpha_i \frac{\mu_1 - \sigma_i}{\mu_1} \right) + \frac{1}{\alpha_i} \frac{\mu_1}{\mu_1 - \sigma_i} - 1.$$

### 4.2.3 Relation with bounded moment models

In MAB literature, Bubeck et al. [2013] showed that MAB models with finite variance are sufficient to achieve the same optimal regret bounds as the MAB problems under sub-Gaussian distributions. Then, several upper confidence bound (UCB) based algorithms were proposed to tackle bandit problems under heavy-tailed noise with infinite variance, under additional assumptions on moments [Lee et al., 2020, Medina and Yang, 2016, Shao et al., 2018].

One major assumption is that the existence of the uniform bound  $\nu_\gamma^*$  satisfying  $\nu_{i,\gamma} \leq \nu_\gamma^*$  for all  $i \in [K]$  is known for some  $\gamma \in [1, \min_{a \in [K]} \gamma_a^*)$  [Agrawal et al., 2021b, Bubeck et al., 2013]. Since the uniform bound of moments is usually unknown in practice, recent research started to assume that only the maximum order  $\gamma^*$  of the finite moment is known [Lee et al., 2020]. Notice that the aforementioned algorithms showed the polynomial problem-dependent regret with assumption on the moments: either known  $\gamma^*$  or known  $\nu_\gamma^*$ . Recall that the  $\gamma$ -th moment of  $\text{Pareto}(\sigma, \alpha)$  is given as

$$\nu_\gamma = \begin{cases} \infty & \alpha \leq \gamma, \\ \frac{\sigma^\gamma \alpha}{\alpha - \gamma} & \alpha > \gamma. \end{cases}$$

Therefore, the assumptions on the moments impose additional restrictions on the parameter space of the Pareto distribution, and the Pareto models and the bounded moment models are not a subset of each other. Recently, Agrawal et al. [2021b] proposed an asymptotically optimal KL-UCB based algorithm that requires solving the optimization problem at every round. Since the bounded moment model only covers certain Pareto distributions in general, the known optimality result of KL-UCB does not necessarily imply the optimality in the sense of (4.2).

### 4.3 Thompson Sampling and the Choice of Priors

In this section, we instantiate TS and a variant of TS, TS-T, for the Pareto model with the choice of noninformative priors under probability matching criteria.

TS is a policy from the Bayesian viewpoint, where the choices of noninformative priors are important as we showed in Chapter 3. Although one can utilize prior knowledge on parameters when choosing the prior, such information would not always be available in practice. To deal with such scenarios, we consider noninformative priors based on the FI matrix, which does not assume any information on unknown parameters.

As discussed in Section 2.3.2, we can obtain the FI matrix if a distribution satisfies the FI regular condition in Definition 2.3. However,  $\text{Pareto}(\sigma, \alpha)$  does not satisfy this condition since it is a parametric-support family. Therefore, for  $X$  with density function in (4.1), one can obtain the FI matrix of  $\text{Pareto}(\sigma, \alpha)$  based on Definition 2.2 as follows [Li et al., 2020]:

$$I(\sigma, \alpha) = \begin{bmatrix} I_{11} & 0 \\ 0 & I_{22} \end{bmatrix} = \begin{bmatrix} \frac{\alpha^2}{\sigma^2} & 0 \\ 0 & \frac{1}{\alpha^2} \end{bmatrix} = \begin{bmatrix} h_{11}(\sigma)h_{11}(\alpha) & 0 \\ 0 & h_{22}(\sigma)h_{22}(\alpha) \end{bmatrix}, \quad (4.4)$$

where  $h_{11}(\sigma) = \frac{1}{\sigma^2}$ ,  $h_{11}(\alpha) = \alpha^2$ ,  $h_{22}(\sigma) = 1$ , and  $h_{22}(\alpha) = \frac{1}{\alpha^2}$ . Note that

$$I_{11} \neq -\mathbb{E} \left[ \frac{\partial^2}{\partial \sigma^2} \log f_{\sigma, \alpha}^{\text{Pareto}}(X; \theta) \middle| \theta \right] = \frac{\alpha}{\sigma^2}.$$

Therefore, the Jeffreys prior is given as

$$\pi_j(\sigma, \alpha) \propto \sqrt{\det(I)} = \frac{1}{\sigma}.$$

Since the FI matrix in (4.4) is diagonal, the reference prior and the probability matching prior when  $\sigma$  is of interest and  $\alpha$  is the nuisance parameter are given as

$$\pi_r(\sigma, \alpha) \propto \sqrt{h_{11}(\sigma)h_{22}(\alpha)} = \frac{1}{\sigma\alpha} \quad \text{and} \quad \pi_{\text{pm}}(\sigma, \alpha) \propto \sqrt{I_{11}}g_1(\alpha) = \frac{\alpha}{\sigma}g_1(\alpha),$$

respectively, for arbitrary  $g_1(\alpha) > 0$  [Tibshirani, 1989]. Here, the reverse reference prior is the same as the reference prior from the orthogonality of parameters [Datta, 1996, Datta and Ghosh, 1995]. In this chapter, we consider the priors

$$\pi_{\text{pm}}^k(\sigma, \alpha) \propto \frac{\alpha^{-k}}{\sigma}$$

for  $k \in \mathbb{Z}$  since the cases  $k = 0, 1$  correspond to the Jeffreys prior and the (reverse) reference prior, respectively.

**Remark 2. Different types of Pareto distributions** The Pareto distribution discussed in this paper is sometimes referred to as the Pareto type 1 distribution [Arnold, 2008]. On the other hand, Kim et al. [2009] derived several noninformative priors for a special case of the Pareto type 2 distribution called the Lomax distribution [Lomax, 1954], which is a shifted Pareto distribution with support beginning at zero. Therefore, the FI matrix of the Lomax distribution can be written using the negative Hessian.

For the Pareto distributions, let us define two minimal jointly sufficient statistics for arm  $i \in [K]$ ,  $T(X_{i,n}) = (x_i^{(1)}, q_{i,n})$  for the parameters  $(\sigma_i, \alpha_i)$  as follows [Li et al., 2020]:

$$x_i^{(1)} := \min_{s \in [n]} x_{i,s}, \quad q_{i,n} := \sum_{s=1}^n \log(r_{i,s}).$$

Note that  $x_i^{(1)}$  denotes the first-order statistic of sampled rewards of arm  $i$ . Then, the maximum likelihood estimators (MLEs) of  $\sigma, \alpha$  for arm  $i$  given  $n$  rewards and their distributions are given as follows [Malik, 1970]:

$$\begin{aligned} \hat{\sigma}_{i,n} &= x_i^{(1)} \sim \text{Pareto}(\sigma_i, n\alpha_i), \\ \hat{\alpha}_{i,n} &= \frac{n}{q_{i,n} - n \log \hat{\sigma}_{i,n}} \sim \text{InvG}(n-1, n\alpha_i), \end{aligned} \quad (4.5)$$

where  $\text{InvG}(s, b)$  denotes the inverse-gamma distribution with shape  $s > 0$  and scale  $b > 0$ . Note that Malik [1970] further showed the stochastic independence of  $\hat{\sigma}_{i,n}$  and  $\hat{\alpha}_{i,n}$ .

#### 4.3.1 TS and TS-T for the Pareto bandits

Based on the MLEs in (4.5), the marginalized posterior distribution of the shape parameter of arm  $i$  for the prior  $\frac{\alpha^{-k}}{\sigma}$  with  $k \in \mathbb{Z}$  is given as

$$\alpha_i \mid \hat{\theta}_{i,n} \sim \text{Erlang}\left(n-k, \frac{n}{\hat{\alpha}_{i,n}}\right), \quad (4.6)$$

where  $\hat{\theta}_{i,n} = (\hat{\sigma}_{i,n}, \hat{\alpha}_{i,n})$  and  $\text{Erlang}(s, \beta)$  denotes the Erlang distribution with shape  $s$  and rate  $\beta$ . Note that we require  $n_0 \geq \max\{2, k+1\}$  initial plays to avoid improper posteriors and MLE of  $\alpha$ . When the shape parameter  $\alpha_i$  is given as  $\beta$ , the cumulative distribution function (CDF) of the conditional posterior of  $\sigma_a$  is given as

$$\mathbb{P}\left[\sigma_i \leq x \mid \hat{\theta}_{i,n}, \alpha_i = \beta\right] = \left(\frac{x}{\hat{\sigma}_{i,n}}\right)^{\beta n}, \quad (4.7)$$

---

**Algorithm 4** TS / TS-T for the Pareto models

---

- 1: **Parameter:**  $k \in \mathbb{Z}$ ,  $n_0 = \max\{2, k + 1\}$ .
  - 2: **Initialization:** Play every arm  $n_0$  times and compute estimators.
  - 3: **for**  $t = n_0K + 1, \dots, T$  **do**
  - 4: **Sample**  $\tilde{\alpha}_a(t) \sim \text{Erlang}\left(N_a(t) - k, \frac{N_a(t)}{\hat{\alpha}_a(N_a(t))}\right)$ . ▷ TS
  - 5:  $\bar{\alpha}_a(N_a(t)) \leftarrow \min(N_a(t), \hat{\alpha}_a(N_a(t)))$ . ▷ Truncated estimator for TS-T
  - 6: **Sample**  $\tilde{\alpha}_a(t) \sim \text{Erlang}\left(N_a(t) - k, \frac{N_a(t)}{\bar{\alpha}_a(N_a(t))}\right)$ . ▷ TS-T
  - 7: **if**  $\{i \in [K] : \tilde{\alpha}_i(t) \leq 1\} \neq \emptyset$  **then**
  - 8:     Play  $i(t) = \arg \min_{i \in [K]} \tilde{\alpha}_i(t)$ .
  - 9: **else**
  - 10:     Sample  $u_i \sim \text{Uniform}(0, 1)$  for every  $i \in [K]$ .
  - 11:      $\tilde{\sigma}_i(t) = \hat{\sigma}_{i, N_i(t)} u_i^{\frac{1}{(N_i(t)\tilde{\alpha}_i(t))}}$ . ▷ Inverse transform sampling with CDF in (4.7)
  - 12:     Play  $i(t) = \arg \max_{a \in [K]} \frac{\tilde{\sigma}_a(t)\tilde{\alpha}_a(t)}{\tilde{\alpha}_a(t)-1} = \arg \max_{a \in [K]} \tilde{\mu}_a(t)$ .
  - 13: **end if**
  - 14: Observe a reward  $x_{i(t), N_{i(t)}(t)+1}$  and update estimators  $\hat{\sigma}_{i(t)}$  and  $\hat{\alpha}_{i(t)}$ .
- 

if  $0 < x \leq \hat{\sigma}_{i,n}$ . Since one can derive the posteriors following the same steps as Sun et al. [2021], the detailed derivation is postponed to Section 4.6.2. At round  $t$ , we denote the sampled scale and shape parameters of arm  $a$  by  $\tilde{\sigma}_i(t)$  and  $\tilde{\alpha}_i(t)$ , respectively, and the corresponding expected reward by  $\tilde{\mu}_i(t) := \mu(\tilde{\sigma}_i(t), \tilde{\alpha}_i(t))$ . Since the expected reward depends on both sampled parameters, we apply the sequential sampling scheme considered in Chapter 3. Specifically, we first sample the shape parameter from the marginalized posterior in (4.6). Then, we sample the scale parameter given the sampled shape parameter from the CDF of the conditional posterior in (4.7) by using inverse transform sampling. TS based on this sequential procedure is illustrated in Algorithm 4

In Theorem 4.3 given in the next section, TS with the reference prior ( $k = 1$ ) and the Jeffreys prior ( $k = 0$ ) turns out to be suboptimal in view of the lower bound in (4.2). Their suboptimality is mainly due to the behavior of the posterior in (4.6) when  $\hat{\alpha}_1(n)$  is overestimated for small  $N_1(t) = n$ . To overcome such issues, we apply the strategy of TS-T to the Pareto bandits, where we replace  $\hat{\alpha}_{i,n}$  with  $\bar{\alpha}_{i,n} := \min(n, \hat{\alpha}_{i,n})$  in (4.6). Note that such a truncation procedure is especially considered in the posterior sampling by (4.6) and (4.7) based on the design idea introduced in Section 1.4.2. We show that TS-T with the reference prior and the Jeffreys prior can achieve the optimal regret bound in Theorem 4.4.

#### 4.3.2 Interpretation of the prior parameter $k$

The Erlang distribution is a special case of the Gamma distribution, where the shape parameter is a positive integer. If a random variable  $X$  follows  $\text{Erlang}(s, \beta)$ , then it has the density of form

$$f_{s,\beta}^{\text{Er}}(x) = \frac{\beta^s}{\Gamma(s)} x^{s-1} e^{-\beta x} \mathbb{1}[x \in \mathbb{R}_+], \quad (4.8)$$

where  $s \in \mathbb{N}$  and  $\beta > 0$  denote the shape and rate parameter, respectively. Then, the CDF evaluated at  $x > 0$  is given as

$$F_{s,\beta}^{\text{Er}}(x) = \frac{\int_0^{\beta x} t^{s-1} e^{-t} dt}{\Gamma(s)} = \frac{\gamma(s, \beta x)}{\Gamma(s)}, \quad (4.9)$$

where  $\gamma(\cdot, \cdot)$  denotes the lower incomplete gamma function. Since  $\gamma(s+1, x) = s\gamma(s, x) - x^s e^{-x}$  holds, one can observe that for any  $x > 0$

$$F_{s,\beta}^{\text{Er}}(x) \geq F_{s+1,\beta}^{\text{Er}}(x). \quad (4.10)$$

From the sampling procedure of TS and TS-T,  $\tilde{\mu}$  depends on  $\tilde{\sigma}$  only when  $\tilde{\alpha} > 1$  holds since  $\tilde{\alpha} \leq 1$  results in  $\mu(\cdot, \tilde{\alpha}) = \infty$ . Therefore, for any  $\beta > 1$  in (4.7),  $\tilde{\sigma}$  will concentrate on  $\hat{\sigma}$  for sufficiently large  $N_a(t) = n$ . Thus,  $\tilde{\mu}$  will be mainly determined by  $\tilde{\alpha}$  and  $\hat{\sigma}$ , where the choice of  $k$  affects the sampling of  $\tilde{\alpha}$  by (4.6). From (4.10), one could see that the probability of sampling small  $\tilde{\alpha}$  increases as shape  $n - k$  decreases. Therefore,  $\tilde{\mu}$  of suboptimal arms would increase as  $k$  increases for the same  $n$ . In other words, the probability of sampling large  $\tilde{\mu}$  becomes large as  $k$  increases. Therefore, TS with large  $k$  becomes a conservative policy that could frequently play currently suboptimal arms. In contrast, priors with small  $k$  yield an optimistic policy that focuses on playing the current best arm.

#### 4.4 Main Theoretical Results

In this section, we provide regret bounds of TS and TS-T with different choices of  $k \in \mathbb{Z}$ . At first, we show the asymptotic optimality of TS for priors  $\pi(\sigma, \alpha) \propto \frac{\alpha^{-k}}{\sigma}$  with  $k \in \mathbb{Z}_{\geq 2}$ .

**Theorem 4.2.** *Assume that arm 1 is the unique optimal arm with a finite mean. For every  $i \in [K]$ , let  $\varepsilon_i = \min \left\{ \frac{\sigma_i}{\alpha_i(\sigma_i+1)}, \frac{\sigma_i \delta_i}{\mu_i(\mu_i+\delta_i-\sigma_i)+\sigma_i \delta_i}, \frac{\sigma_i \delta_i}{\mu_i(1+\mu_i+\delta_i)} \right\}$  where  $\delta_i = \frac{\Delta_i}{2}$  for  $i \neq 1$  and  $\delta_1 = \min_{i \neq 1} \delta_i$ . Given arbitrary  $\epsilon \in (0, \min_{a \in [K]} \varepsilon_a)$ , the expected regret of TS with  $k \in \mathbb{Z}_{\geq 2}$  is bounded as*

$$\mathbb{E}[\text{Reg}(T)] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{D_{i,k}(\epsilon)} + \mathcal{O}(\epsilon^{-2}),$$

where for  $b_{i,k}(\epsilon) = (1 + (\max(0, k) + 1)\alpha_i \epsilon)^{-1}$ ,

$$D_{i,k}(\epsilon) = \inf_{\theta: \mu(\theta) > \mu_1 - \epsilon} \text{KL}(\text{Pareto}(\sigma_i + \epsilon, \alpha_i b_{i,k}(\epsilon)), \text{Pareto}(\theta)) \quad (4.11)$$

is a function such that  $\lim_{\epsilon \rightarrow 0} D_{i,k}(\epsilon) = \text{KL}_{\text{inf}}(i)$  for any fixed  $k \in \mathbb{Z}$ .

By letting  $\epsilon = o(1)$  in Theorem 4.2, we see that TS with  $k \in \mathbb{Z}_{\geq 2}$  satisfies

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \leq \sum_{i=2}^K \frac{\Delta_i}{\text{KL}_{\text{inf}}(i)},$$

which shows the asymptotic optimality of TS in terms of the lower bound in (4.2).

Next, we show that TS with  $k \in \mathbb{Z}_{\leq 1}$  cannot achieve the asymptotic bound in the theorem below. Similarly to Theorem 3.2, we consider two-armed bandit problems where the full information on the suboptimal arm is given to simplify the analysis.

**Theorem 4.3.** *Assume that the arm 1 follows  $\text{Pareto}(\sigma_1, \alpha_1)$  and the arm 2 follows  $\text{Pareto}(\sigma_2, \alpha_2)$  with  $\sigma_1 < \sigma_2$  and  $1 < \alpha_1 < \alpha_2$ . When  $\tilde{\alpha}_1(t)$  and  $\tilde{\sigma}_1(t)$  are sampled based on the posteriors in (4.6) and (4.7) with prior  $k \in \mathbb{Z}_{\leq 1}$ , respectively and  $\tilde{\mu}_2(t) = \mu_2$  holds, there exists a constant  $\xi^{\text{Pa}} > 0$  independent of  $\alpha_2$  satisfying*

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \geq \Delta_2 \xi^{\text{Pa}},$$



where  $\xi^{\text{Pa}} > \text{KL}_{\text{inf}}(2)$  holds for some instances. In particular, for  $k \in \mathbb{Z}_{\leq 0}$ , there exists a constant  $\xi^{\text{Pa}'} > 0$  independent of  $\alpha_2$  satisfying

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\sqrt{T}} \geq \Delta_2 \xi^{\text{Pa}'}.$$

From Theorems 4.2 and 4.3, we find that the prior should be conservative to some extent when one considers maximizing rewards in *expectation*.

Although TS with the Jeffreys prior ( $k = 0$ ) and reference prior ( $k = 1$ ) were shown to be suboptimal, we show that TS-T can achieve the optimal regret bound with  $k \in \mathbb{Z}_{\geq 0}$ .

**Theorem 4.4.** *With the same notation as Theorem 4.2, the expected regret of TS-T with  $k \in \mathbb{Z}_{\geq 0}$  is bounded as*

$$\mathbb{E}[\text{Reg}(T)] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{D_{i,k}(\epsilon)} + \mathcal{O}(\epsilon^{-m}),$$

where  $m = \max(2, 3 - k)$ .

From Theorems 4.2 and 4.4, we have two choices to achieve the lower bound in (4.2): use either the conservative priors with MLEs or moderately optimistic priors with truncated samples. Since initialization steps require playing every arm  $\max(2, k + 1)$  times, if the number of arms  $K$  is large, the reference priors or the Jeffreys prior with the truncated estimator would be a better choice. On the other hand, if the model can contain arms with large  $\alpha$ , where the truncation might be problematic for small  $n$ , it would be better to use TS with conservative priors.

**Remark 3. The uniform prior** The uniform prior with  $(\sigma, \alpha)$  parameterization, i.e.,  $\pi_{\text{u}}(\sigma, \alpha) \propto 1$  cannot be represented in the probability matching priors considered in this chapter,  $\pi_{\text{pm}}(\sigma, \alpha)$ . One reason to choose  $\pi_{\text{pm}}$  is the simplicity of the implementation as we can obtain the closed form of the posterior, which preserves one of the main advantages of TS. On the other hand, the marginalized posterior density of  $\alpha$  based on  $\pi_{\text{u}}(\sigma, \alpha)$  can be approximated as

$$\pi_{\text{u}}(\alpha \mid T(X_{i,n})) \propto \frac{\alpha^n}{n\alpha + 1} \exp\left(-\frac{n}{\hat{\alpha}_{i,n}}\alpha\right),$$

which cannot be written as some well-known distributions. In contrast, the posterior density based on  $\pi_{\text{pm}}(\sigma, \alpha)$  is written as

$$\pi_{\text{pm}}^k(\alpha \mid T(X_{i,n})) \propto \alpha^{n-k-1} \exp\left(-\frac{n}{\hat{\alpha}_{i,n}}\alpha\right),$$

which is a density function of the Erlang distribution in (4.8). Based on the above formulations, we expect that the uniform prior with  $(\sigma, \alpha)$  parameterization will behave similarly to the probability matching priors with  $k \in (0, 1)$ . In other words, the uniform prior with  $(\sigma, \alpha)$  parameterization is expected to be more optimistic than the reference prior and more conservative than the Jeffreys prior.

Similar to the location-scale family in Chapter 3, one can consider the uniform prior on the rate-shape parameterization, which is  $\pi_{\text{u}}(\sigma^{-1}, \alpha) \propto 1$ . However, by following the same steps in Corollary 3.3, one can find that  $\pi_{\text{u}}(\sigma^{-1}, \alpha)$  is equivalent to the Jeffreys prior on the scale-shape parameterization. In summary, the uniform prior is not only difficult to implement but also suboptimal for the Pareto bandits.

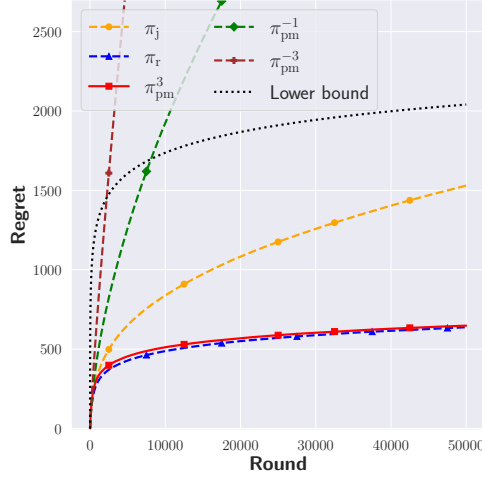
Table 4.1: Parameters of the 4-armed bandit instances.

(a) Instance  $\nu_4^{(1)}$  where suboptimal arms have smaller, equal, and larger  $\sigma$  compared with the optimal arm.

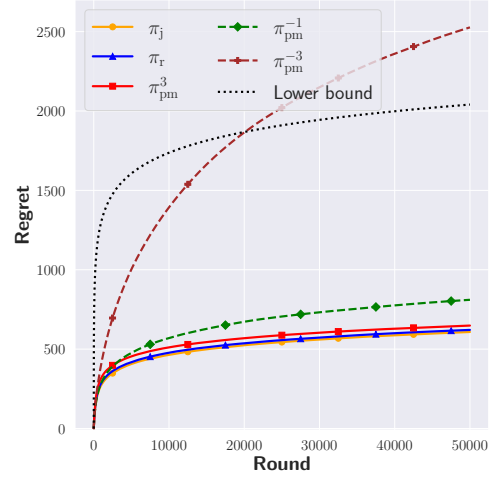
	Arm 1	Arm 2	Arm 3	Arm 4
$\sigma$	1.3	1.2	1.3	1.5
$\alpha$	1.4	1.6	1.9	2.0
$\mu$	4.55	3.2	2.74	3.0

(b) Challenging instance  $\nu_4^{(2)}$  where some suboptimal arms have twice larger  $\sigma$  than the optimal arm.

	Arm 1	Arm 2	Arm 3	Arm 4
$\sigma$	1.0	1.5	2.0	2.0
$\alpha$	1.2	1.5	1.8	2.0
$\mu$	5.0	4.5	4.5	4.0



(a) Regret of TS.



(b) Regret of TS-T.

Figure 4.1: Cumulative regret for the 4-armed Pareto bandit instance  $\nu_4^{(1)}$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively. The black dotted line denotes the problem-dependent lower bound based on Lemma 4.1.

## 4.5 Simulation Results

In this section, we present numerical results to demonstrate the performance of TS and TS-T, which supports our theoretical analysis. We consider two different 4-armed bandit models  $\nu_4^{(1)}$  and  $\nu_4^{(2)}$  with parameters given in Table 4.1. The first instance  $\nu_4^{(1)}$  is an example where suboptimal arms have smaller, equal, and larger  $\sigma$  compared with the optimal arm. The second instance  $\nu_4^{(2)}$  would be a more challenging problem than  $\nu_4^{(1)}$  in the sense that the  $\sigma$  determines the left boundary of the support, where larger  $\sigma$  implies a larger minimum value of the arm. Therefore, if  $\sigma$  of the suboptimal arm is larger than that of the optimal arm, it would make a problem difficult in the first few trials.

Figures 4.1 and 4.3 show the cumulative regret for the proposed policies with various choices of parameters  $k$  on the prior. The solid lines denote the averaged cumulative regret over 10,000 independent runs of priors that can achieve the optimal lower bound in (4.2), whereas the dashed lines denote that of priors that cannot. The green dotted line denotes the problem-dependent lower bound and shaded regions denote a quarter standard deviation.

In Figures 4.2 and 4.4, we investigate the difference between TS and TS-T with the same prior parameter  $k$ . The solid lines denote the averaged cumulative regret over 10,000 independent runs. The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of regret.

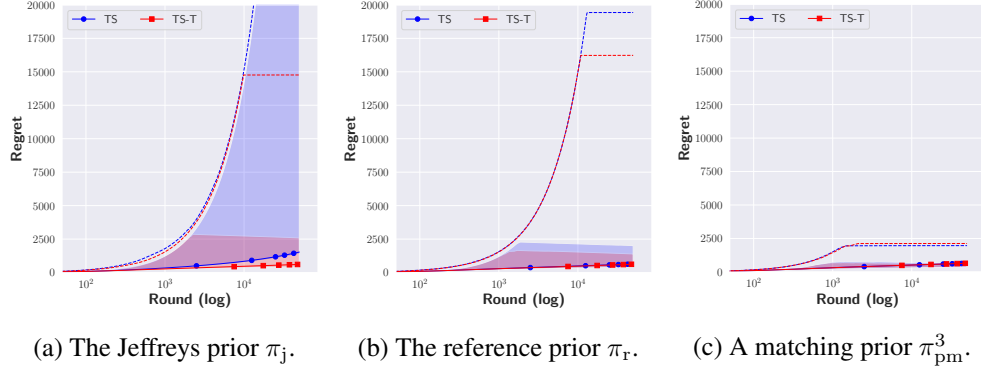


Figure 4.2: The solid lines denote an averaged regret over independent 10,000 runs for the 4-armed Pareto bandit instance  $\nu_4^{(1)}$ . The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of the regret, respectively.

**The Jeffreys prior  $\pi_j$  ( $\pi_{pm}^0$ )** In Figures 4.1a and 4.3a, the Jeffreys prior seems to have a larger order of the regret compared with priors  $\pi_r$  and  $\pi_{pm}^3$ , which performed the best in this setting. As Theorem 4.4 states, its performance improves under TS-T, which shows a similar performance to that of  $\pi_r$  and  $\pi_{pm}^3$ .

Figures 4.2a and 4.4a illustrate the possible reason for the improvements, where the central 99% interval of the regret noticeably shrank under TS-T. Since the suboptimality of TS with the Jeffreys prior ( $k = 0$ ) is due to an extreme case that induces a polynomial regret with a small probability, this kind of shrink contributes to decreasing the expected regret of TS-T with the Jeffreys prior.

**The reference prior  $\pi_r$  ( $\pi_{pm}^1$ )** The reference prior showed a similar performance to the asymptotically optimal prior  $k = 3$ , although it was shown to be suboptimal for some instances under TS in Theorem 4.3. Similarly to the Jeffreys prior ( $k = 0$ ), the reference prior ( $k = 1$ ) under TS-T has a smaller central 99% interval of the regret than that under TS as shown in Figures 4.2b and 4.4b, although its decrement is comparably smaller than that of the Jeffreys prior. This would imply that the reference prior is more conservative than the Jeffreys prior.

**The conservative prior ( $\pi_{pm}^3$ )** Interestingly, Figures 4.2c and 4.4c showed that a truncated procedure does not affect the central 99% interval of the regret and even degrade the performance in upper 0.05%. Notice that the upper 0.05% of the regret of  $k = 3$  is much lower than that of  $k = 0, 1$ , which shows the stability of the conservative prior in Figure 4.2.

Since a truncation procedure was adopted to prevent an extreme case that was a problem for  $k \in \mathbb{Z}_{\leq 1}$ , it is natural to see that there is no difference between TS and TS-T with  $k = 3$ . This would imply that  $k = 3$  is sufficiently conservative, and so the truncated procedure does not affect the overall performance.

**Optimistic priors ( $\pi_{pm}^{-1}$  and  $\pi_{pm}^{-3}$ )** In Figures 4.1a and 4.3a, one can see that the averaged regret of  $k = -1$  and  $k = -3$  increases much faster than that of  $k = 0, 1, 3$  under the TS policy, which illustrates the suboptimality of TS with priors  $k \in \mathbb{Z}_{<0}$ .

As the optimistic priors ( $k < 0$ ) showed better performance under TS-T in Figures 4.1 and 4.3, we can check the effectiveness of a truncation procedure in the posterior sampling with optimistic priors. However, one can also observe that TS-T with these optimistic priors does not perform well compared to that with other optimal priors, which

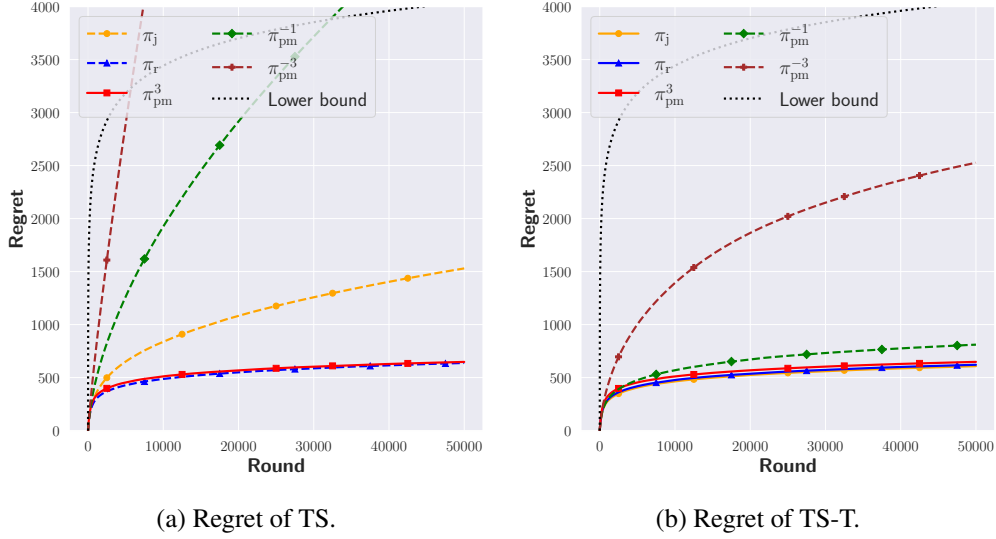


Figure 4.3: Cumulative regret for the 4-armed Pareto bandit instance  $\nu_4^{(2)}$ . The solid lines and the dashed lines denote the averaged values over 10,000 independent runs of the policies that can and cannot achieve the regret lower bound, respectively. The green dotted line denotes the problem-dependent lower bound based on Lemma 4.1.

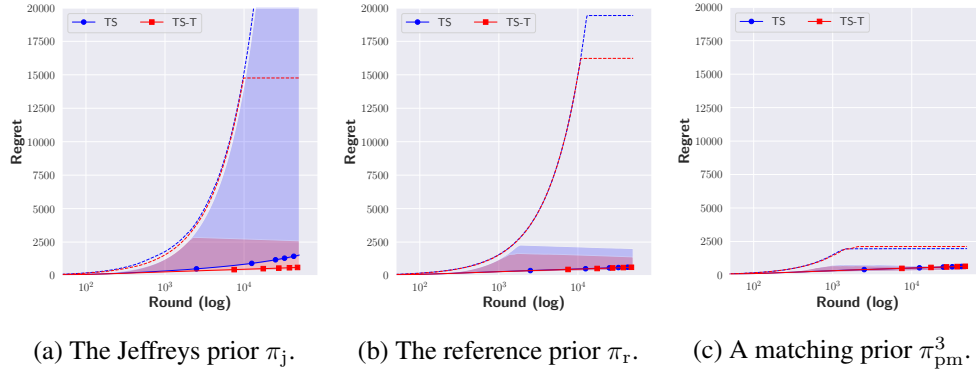


Figure 4.4: The solid lines denote an averaged regret over independent 10,000 runs for the 4-armed Pareto bandit instance  $\nu_4^{(2)}$ . The shaded regions and dashed lines show the central 99% interval and the upper 0.05% of the regret, respectively.

might imply that a prior with  $k \in \mathbb{Z}_{<0}$  is too optimistic. Therefore, we might need to use a more conservative truncation procedure such as the one using  $\bar{\alpha}_{a,n} = \max(\sqrt{n}, \hat{\alpha}_{a,n})$  or  $\max(\log n, \hat{\alpha}_{a,n})$ , which would induce a larger regret in the finite time horizon.

## 4.6 Proofs of Theoretical Results

This section provides detailed proofs for all the theorems and lemmas presented in this chapter.

### 4.6.1 Closed form of the problem-dependent constant

Here, we provide the proof of Lemma 4.1, which derives the closed form of the problem-dependent constant that illustrates the minimum discrimination information for the hypothesis that the arm  $i$  is optimal when the optimal arm has the mean  $\mu_1$ .

*Proof.* Recall the definition

$$\text{KL}_{\inf}(i) = \text{KL}_{\inf}(\text{Pareto}(\theta_i), \text{Pareto}(\theta)) := \inf_{\theta \in \Theta_i} \log \frac{\alpha_i}{\alpha} + \alpha \log \frac{\sigma_i}{\sigma} + \frac{\alpha}{\alpha_i} - 1,$$

where  $\theta = (\sigma, \alpha)$  and  $\Theta_i$  defined in (4.3). Here, we consider the partition of  $\Theta_i$ ,

$$\begin{aligned} \Theta_i^{(1)} &= \{(\sigma, \alpha) \in (0, \sigma_i] \times (0, 1] : \mu(\sigma, \alpha) > \mu_1\} = (0, \sigma_i] \times (0, 1] \\ \Theta_i^{(2)} &= \left\{(\sigma, \alpha) \in (0, \sigma_i] \times (1, \infty) : \mu(\sigma, \alpha) = \frac{\sigma\alpha}{\alpha-1} > \mu_1\right\}, \end{aligned} \quad (4.12)$$

where  $\Theta_i^{(1)} \cup \Theta_i^{(2)} = \Theta_i$ . Therefore, it holds that

$$\begin{aligned} \text{KL}_{\inf}(i) &= \min \left( \inf_{\theta \in \Theta_i^{(1)}} \log \frac{\alpha_i}{\alpha} + \alpha \log \frac{\sigma_i}{\sigma} + \frac{\alpha}{\alpha_i} - 1 \right. \\ &\quad \left. , \inf_{\theta \in \Theta_i^{(2)}} \log \frac{\alpha_i}{\alpha} + \alpha \log \frac{\sigma_i}{\sigma} + \frac{\alpha}{\alpha_i} - 1 \right). \end{aligned}$$

**Case of  $\Theta_i^{(1)}$**

For  $(\sigma, \alpha) \in \Theta_i^{(1)}$ ,  $\mu(\sigma, \alpha) = \infty$  holds regardless of  $\sigma$ . Therefore, we obtain

$$\begin{aligned} \inf_{\theta \in \Theta_i^{(1)}} \log \frac{\alpha_i}{\alpha} + \alpha \log \frac{\sigma_i}{\sigma} + \frac{\alpha}{\alpha_i} - 1 &= \inf_{\alpha \in (0, 1]} \log \frac{\alpha_i}{\alpha} + \frac{\alpha}{\alpha_i} - 1 \\ &= \log \alpha_i + \frac{1}{\alpha_i} - 1, \end{aligned}$$

where the last equality holds since  $\log x + \frac{1}{x} - 1$  is an increasing function for  $x \geq 1$ .

**Case of  $\Theta_i^{(2)}$**

Let  $c = \frac{\sigma_i}{\sigma} \geq 1$  to make KL divergence from  $\text{Pareto}(\theta_i)$  to  $\text{Pareto}(\sigma, \alpha)$  be well-defined. From its definition of  $\Theta_i^{(2)}$  in (4.12), any  $\theta = (\sigma, \alpha) \in \Theta_i^{(2)}$  satisfies  $\frac{\sigma\alpha}{\alpha-1} \geq \mu_1$ , i.e.,

$$\frac{\sigma_i \alpha}{c(\alpha-1)} \geq \mu_1 \Leftrightarrow \alpha \leq \frac{c\mu_1}{c\mu_1 - \sigma_i} =: h_i(c, \boldsymbol{\mu}) = h_i(c).$$

Note that it holds that

$$h_i(1) = \frac{\mu_1}{\mu_1 - \sigma_i} \leq \frac{\mu_i}{\mu_i - \sigma_i} = \alpha_i$$

since  $\frac{x}{x-y}$  is decreasing with respect to  $x \geq y$ . Then, we can rewrite the infimum of KL divergence for  $\theta \in \Theta_i^{(2)}$  as

$$\inf_{\theta \in \Theta_i^{(2)}} \log \frac{\alpha_i}{\alpha} + \alpha \log \frac{\sigma_i}{\sigma} + \frac{\alpha}{\alpha_i} - 1 = \inf_{c \geq 1} \inf_{\alpha \leq h_i(c)} \log \frac{\alpha_i}{\alpha} + \alpha \log c + \frac{\alpha}{\alpha_i} - 1.$$

Define the function in RHS as  $g_i(\alpha, c) := \log \frac{\alpha_i}{\alpha} + \alpha \log c + \frac{\alpha}{\alpha_i} - 1$  that satisfies

$$\frac{\partial g_i(\alpha, c)}{\partial \alpha} = \frac{1}{\alpha_i} + \log c - \frac{1}{\alpha}.$$

Therefore, the infimum value of the inner function can be obtained when  $\alpha = \frac{\alpha_i}{1 + \alpha_i \log c}$  if  $\frac{\alpha_i}{1 + \alpha_i \log c} < h_i(c)$ , where the infimum value is  $g_i\left(\frac{\alpha_i}{1 + \alpha_i \log c}, c\right) = \log(1 + \alpha_i \log c)$ .

Let  $c_i^* \geq 1$  be a deterministic constant such that

$$\begin{aligned} h_i(c_i^*) &= \frac{c_i^* \mu_1}{c_i^* \mu_1 - \sigma_i} = \frac{\alpha_i}{1 + \alpha_i \log c_i^*} \\ \iff (\mu_1 \alpha_i) c_i^* \log c_i^* + (\mu_1 - \alpha_i \mu_1) c_i^* &= -\alpha_i \sigma_i \end{aligned} \quad (4.13)$$

so that  $h_i(c) \geq \frac{\alpha_i}{1 + \alpha_i \log c}$  holds for any  $c \geq c_i^*$ . Since the solution of  $ax \log(x) + bx = -c$  is  $\exp\left(W\left(-\frac{ce^{b/a}}{a}\right) - \frac{b}{a}\right)$  for the principal branch of Lambert W function  $W(\cdot)$ , one can obtain  $c_i^*$  by solving the equality in (4.13), which is

$$c_i^* = \exp\left(W\left(-\frac{\sigma_i}{\mu_1} e^{\frac{1}{\alpha_i}-1}\right) + 1 - \frac{1}{\alpha_i}\right). \quad (4.14)$$

Notice that  $\frac{\sigma_i}{\mu_1} e^{\frac{1}{\alpha_i}-1} \leq \frac{\sigma_i}{\mu_i} e^{\frac{1}{\alpha_i}-1} \leq \left(1 - \frac{1}{\alpha_i}\right) e^{-\left(1-\frac{1}{\alpha_i}\right)} \leq e^{-1}$  holds so that  $c_i^*$  is a real value. Here, we consider the principal branch to ensure  $c_i^* \geq 1$  since the solution on other branches,  $W_{-1}(\cdot)$ , is less than 1, which is out of our interest.

Let  $A_i = 1 - \frac{1}{\alpha_i}$ , which is positive as  $\alpha_i > 1$ , and  $B_i = \frac{\sigma_i}{\mu_1}$ . Then, we can rewrite  $c_i^*$  as

$$c_i^* = e^{A_i} e^{W(-B_i e^{-A_i})} = e^{A_i} e^{-A_i} \frac{-B_i}{W(-B_i e^{-A_i})}. \quad \because e^{W(x)} = \frac{x}{W(x)}$$

Since  $B_i \leq \frac{\sigma_i}{\mu_i} = \frac{\alpha_i - 1}{\alpha_i} = A_i$  holds and the principal branch of Lambert W function is increasing for  $x \geq -\frac{1}{e}$ , we have

$$0 > W(-B_i e^{-A_i}) \geq W(-B_i e^{-B_i}) = -B_i,$$

which implies that  $c_i^* \geq 1$ . Therefore, the infimum of  $g_i$  can be written as

$$\begin{aligned} \inf_{c \geq 1} \inf_{\alpha \leq h_i(c)} g_i(\alpha, c) &= \min\left(\inf_{c \in [1, c_i^*]} g_i(h_i(c), c), \inf_{c \geq c_i^*} \log(1 + \alpha_i \log c)\right) \\ &= \min\left(\inf_{c \in [1, c_i^*]} g_i(h_i(c), c), \log(1 + \alpha_i \log c_i^*)\right), \end{aligned}$$

where we follow the convention that the infimum over the empty set is defined as infinity.

By substituting  $c_i^*$  in (4.14), we obtain

$$\log(1 + \alpha_i \log c_i^*) = \log\left(\alpha_i + W\left(-\frac{\sigma_i}{\mu_1} e^{\frac{1}{\alpha_i}-1}\right)\right).$$

Let us consider the following inequalities:

$$\begin{aligned} \log\left(\alpha_i + W\left(-\frac{\sigma_i}{\mu_1} e^{\frac{1}{\alpha_i}-1}\right)\right) &\geq \log\left(\alpha_i + W\left(-\frac{\sigma_i}{\mu_i} e^{\frac{1}{\alpha_i}-1}\right)\right) \\ &= \log\left(\alpha_i + W\left(\frac{1 - \alpha_i}{\alpha_i} e^{\frac{1}{\alpha_i}-1}\right)\right) \\ &= \log\left(\alpha_i + \frac{1}{\alpha_i} - 1\right), \end{aligned} \quad (4.15)$$

where the first inequality holds since the principal branch of Lambert W function  $W(x)$  is increasing and negative with respect to  $x \in [-1/e, 0)$ .

It remains to find the closed form of  $\inf_{c \in [1, c_i^*]} g_i(h_i(c), c)$ . From the definition of  $g_i(x, c) = \log \frac{\alpha_i}{x} + x \log c + \frac{x}{\alpha_i} - 1$  and  $h_i(c) = \frac{c\mu_1}{c\mu_1 - \sigma_i}$ , we have  $h_i'(c) = -\frac{\mu_1\sigma_i}{(c\mu_1 - \sigma_i)^2}$  and

$$\begin{aligned} \frac{\partial g_i(h_i(c), c)}{\partial c} &= \frac{\partial}{\partial c} \left( \log \frac{\alpha_i}{h_i(c)} + h_i(c) \log c + \frac{h_i(c)}{\alpha_i} - 1 \right) \\ &= -\frac{h_i'(c)}{h_i(c)} + h_i'(c) \log c + \frac{h_i(c)}{c} + \frac{1}{\alpha_i} h_i'(c) \\ &= \frac{\sigma_i}{c(c\mu_1 - \sigma_i)} - \frac{\mu_1\sigma_i}{(c\mu_1 - \sigma_i)^2} \log c + \frac{\mu_1}{c\mu_1 - \sigma_i} - \frac{\sigma_i\mu_1}{\alpha_i(c\mu_1 - \sigma_i)^2} \\ &= \frac{\sigma_i}{c(c\mu_1 - \sigma_i)} - \frac{\mu_1\sigma_i}{(c\mu_1 - \sigma_i)^2} \log c + \mu_1 \frac{c\mu_1 - \sigma_i - \frac{\sigma_i}{\alpha_i}}{(c\mu_1 - \sigma_i)^2}. \end{aligned} \quad (4.16)$$

Since the first term in (4.16) is positive for  $c \geq 1$  and  $\mu_1 \geq \mu_i > \sigma_i$ , let us consider the last two terms for  $c \in [1, c_i^*]$ ,

$$\begin{aligned} -\frac{\mu_1\sigma_i}{(c\mu_1 - \sigma_i)^2} \log c + \mu_1 \frac{c\mu_1 - \sigma_i - \frac{\sigma_i}{\alpha_i}}{(c\mu_1 - \sigma_i)^2} \\ &= \frac{\mu_1}{(c\mu_1 - \sigma_i)^2} \left( c\mu_1 - \sigma_i - \frac{\sigma_i}{\alpha_i} - \sigma_i \log c \right) \\ &= \frac{\mu_1}{(c\mu_1 - \sigma_i)^2} \left( \mu_1 - \sigma_i - \frac{\sigma_i}{\alpha_i} + (c-1)\mu_1 - \sigma_i \log c \right) \\ &= \frac{\mu_1}{(c\mu_1 - \sigma_i)^2} \left( \mu_1 - \sigma_i - \frac{\sigma_i}{\alpha_i} + \mu_1 \left( c - \frac{\sigma_i}{\mu_1} \log c - 1 \right) \right). \end{aligned}$$

Here,

$$\mu_1 - \sigma_i - \frac{\sigma_i}{\alpha_i} \geq \mu_i - \sigma_i - \frac{\sigma_i}{\alpha_i} = \frac{\sigma_i\alpha_i}{\alpha_i - 1} - \sigma_i - \frac{\sigma_i}{\alpha_i} = \frac{\sigma_i}{\alpha_i(\alpha_i - 1)} > 0,$$

and  $c - \frac{\sigma_i}{\mu_1} \log c - 1$  is increasing with respect to  $c$  so that  $c - \frac{\sigma_i}{\mu_1} \log c - 1 \geq 0$  for  $c \geq 1$ . Therefore,  $\frac{\partial}{\partial c} g_i(h_i(c), c)$  is positive for  $c \geq 1$ , i.e.,  $g_i(h_i(c), c)$  is an increasing function with respect to  $c \geq 1$ . Thus, we have for  $c \in [1, c_i^*]$ ,

$$\begin{aligned} \inf_{c \in [1, c_i^*]} g_i(h_i(c), c) &= g_i(h_i(1), 1) = g_i\left(\frac{\mu_1}{\mu_1 - \sigma_i}, 1\right) \\ &= \log \left( \alpha_i \frac{\mu_1 - \sigma_i}{\mu_1} \right) + \frac{1}{\alpha_i} \frac{\mu_1}{\mu_1 - \sigma_i} - 1. \end{aligned}$$

## Summary

Based on the above results, we have

$$\text{KL}_{\inf}(i) = \min \left( \log \alpha_i + \frac{1}{\alpha_i} - 1, \inf_{c \in [1, c_i^*]} g_i(h_i(c), c), \log(1 + \alpha_i \log c_i^*) \right)$$

Notice that  $\alpha_i \frac{\mu_1 - \sigma_i}{\mu_1} \in [1, \alpha_i]$  where the equality occurs only at  $\mu_i = \mu_1$ . Since  $\log x + \frac{1}{x} - 1$  is increasing with respect to  $x \geq 1$ , we have for  $\alpha_i > 1$

$$\begin{aligned} \log \left( \alpha_i \frac{\mu_1 - \sigma_i}{\mu_1} \right) + \frac{1}{\alpha_i} \frac{\mu_1}{\mu_1 - \sigma_i} - 1 &\leq \log \alpha_i + \frac{1}{\alpha_i} - 1 \\ &\leq \log \left( \alpha_i + \frac{1}{\alpha_i} - 1 \right) \\ &\leq \log(1 + \alpha_i \log c_i^*), \end{aligned}$$

where the last inequality comes from the result in (4.15). Therefore, we have

$$\text{KL}_{\text{inf}}(i) = \log \left( \alpha_i \frac{\mu_1 - \sigma_i}{\mu_1} \right) + \frac{1}{\alpha_i} \frac{\mu_1}{\mu_1 - \sigma_i} - 1,$$

which concludes the proof.  $\square$

#### 4.6.2 Derivation of the posteriors

In this section, we provide the detailed derivation of the posteriors based on the probability matching prior  $\frac{\alpha^{-k}}{\sigma}$ .

Let the observation  $X_n = (x_1, \dots, x_n)$  of an arm and let  $q_n = \sum_{s=1}^n \log x_s$ . Then, Bayes' theorem gives the posterior density as

$$p(\sigma, \alpha \mid X_n) = \frac{p(X_n \mid \sigma, \alpha) p(\sigma, \alpha)}{\int_0^\infty \int_0^\infty p(X_n \mid \sigma, \alpha) p(\sigma, \alpha) d\sigma d\alpha},$$

where

$$\begin{aligned} p(X_n \mid \sigma, \alpha) &= \alpha^n \sigma^{n\alpha} \left( \prod_{s=1}^n x_s \right)^{-\alpha-1} \mathbb{1}[\sigma \leq \hat{\sigma}_n] \\ &= \alpha^n \sigma^{n\alpha} \exp(-q_n(\alpha + 1)) \mathbb{1}[\sigma \leq \hat{\sigma}_n]. \end{aligned}$$

By performing a direct computation, we obtain for  $k \in \mathbb{Z}$  that

$$\begin{aligned} \int_0^\infty \int_0^\infty p(X_n \mid \sigma, \alpha) \pi_{\text{pm}}^k(\sigma, \alpha) d\sigma d\alpha &= \int_0^\infty \int_0^\infty p(X_n \mid \sigma, \alpha) \frac{\alpha^{-k}}{\sigma} d\sigma d\alpha \\ &= \int_0^\infty \alpha^{n-k} \exp(-q_n(\alpha + 1)) \int_0^{\hat{\sigma}} \sigma^{n\alpha-1} d\sigma d\alpha \\ &= \int_0^\infty \frac{\alpha^{n-k-1}}{n} e^{-q_n} \exp(-\alpha(q_n - n \log \hat{\sigma})) d\alpha \\ &= \frac{\Gamma(n-k)}{n} \frac{e^{-q_n}}{(q_n - n \log \hat{\sigma})^{n-k}}. \end{aligned}$$

Therefore, the joint posterior probability density is given as follows:

$$p(\sigma, \alpha \mid X_n) = \frac{n[q_n - n \log \hat{\sigma}_n]^{n-k}}{\Gamma(n-k)} \alpha^{n-k} \sigma^{n\alpha-1} e^{-q_n \alpha} \mathbb{1}[0 < \sigma \leq \hat{\sigma}_n],$$

which gives the marginal posterior of  $\alpha$  as

$$\pi^k(\alpha \mid X_n) = \frac{\alpha^{n-k-1} [q_n - n \log \hat{\sigma}_n]^{n-k}}{\Gamma(n-k)} e^{-\alpha(q_n - n \log \hat{\sigma}_n)}.$$

Thus, sample  $\tilde{\alpha}$  generated from the marginal posterior actually follows the Gamma distribution with shape  $n-k$  and rate  $q_n - n \log \hat{\sigma}_n = \frac{n}{\tilde{\alpha}}$ , i.e.,  $\tilde{\alpha} \sim \text{Erlang}(n-k, \frac{n}{\tilde{\alpha}})$  as  $n \in \mathbb{N}$  and  $k \in \mathbb{Z}$  if  $n > k$ . When  $\tilde{\alpha}$  is given, the conditional posterior of  $\sigma$  is given as

$$\begin{aligned} p(\sigma \mid X_n, \alpha) &= \frac{p(\sigma, \alpha \mid X_n)}{p(\alpha \mid X_n)} \\ &= \frac{n\alpha}{\hat{\sigma}^{n\alpha}} \sigma^{n\alpha-1} \mathbb{1}[0 < \sigma \leq \hat{\sigma}_n]. \end{aligned}$$

Hence, the cumulative distribution function (CDF) of  $\sigma$  given  $\alpha$  is given as

$$\mathbb{P}(\sigma \leq x) = F(x \mid X_n, \alpha = \tilde{\alpha}) = \left( \frac{x}{\hat{\sigma}_n} \right)^{n\tilde{\alpha}}, \quad 0 < x \leq \hat{\sigma}_n.$$



Note that MLEs of  $\sigma, \alpha$  are equivalent to the maximum a posteriori (MAP) estimators when one uses the Jeffreys prior [Li et al., 2020, Sun et al., 2021].

In sum, based on  $\pi_{\text{pm}}^k$ , we have the marginalized posterior distribution of  $\alpha$

$$\alpha \mid X_n \sim \text{Erlang}\left(n - k, \frac{n}{\hat{\alpha}}\right)$$

and the cumulative distribution function (CDF) of the conditional posterior of  $\sigma$

$$F(x \mid X_n, \alpha = \tilde{\alpha}) = \left(\frac{x}{\hat{\alpha}_n}\right)^{n\tilde{\alpha}}, \quad 0 < x \leq \hat{\sigma}_n.$$

Note that we require  $\max\{2, k + 1\}$  initial plays to avoid improper posteriors and improper MLEs.

### 4.6.3 Proof of the optimality of TS and TS-T

In this section, we provide the proof of Theorems 4.2 and 4.3, whose overall structure is similar to that in the previous chapter.

*Proof.* Recall the regret decomposition in (3.9), which shows

$$\begin{aligned} \text{Reg}(T) &= \sum_{i=2}^K \Delta_i n_0 + \underbrace{\sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1}[i(t) = i, \mathcal{M}_\epsilon^c(t)]}_{\text{bad optimal (BO) term}} \\ &\quad + \underbrace{\sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1}[i(t) = i, \mathcal{M}_\epsilon(t)]}_{\text{good optimal (GO) term}}, \end{aligned}$$

where  $\mathcal{M}_\epsilon(t) = \{\tilde{\mu}^*(t) \geq \mu_1 - \epsilon\}$ . Then, the following lemmas below conclude the proof of Theorems 4.2 and 4.3.

**Lemma 4.5.** *For the  $K$ -armed Pareto bandit models, it holds under both TS and TS-T that*

$$\mathbb{E}[(\text{GO})] \leq \sum_{i=2}^K \frac{\Delta_i \log T}{D_{i,k}(\epsilon)} + \mathcal{O}(k) + \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-2}).$$

Since large  $k$  yields a more conservative policy and requires additional initial plays of every arm, large  $k$  might induce larger regret for a finite time horizon  $T$ , which corresponds to the component of the regret discussed in Lemma 4.7. Thus, this lemma would imply that the policy has to be conservative to some extent, and being overly conservative would induce larger regrets in a finite time.

**Lemma 4.6.** *For the  $K$ -armed Pareto bandit models, it holds that*

$$\mathbb{E}[(\text{BO})] \leq \begin{cases} \mathcal{O}(\epsilon^{-2}) & \text{under TS with prior } \frac{\alpha^{-k}}{\sigma}, k \in \mathbb{Z}_{\geq 2}, \\ \mathcal{O}(\epsilon^{-\max(2, 3-k)}) & \text{under TS-T with prior } \frac{\alpha^{-k}}{\sigma}, k \in \mathbb{Z}_{\geq 0}. \end{cases}$$

### 4.6.4 Proof of Lemma 4.5

In this section, we provide the upper bound of the (GO) term that contains the main regret term.

Let us first consider good events on MLEs defined by

$$\begin{aligned}\mathcal{A}_{i,n}(\epsilon) &:= \{\hat{\alpha}_{i,n} \in [\alpha_i - \epsilon_{i,l}(\epsilon), \alpha_i + \epsilon_{i,u}(\epsilon)]\} \\ \mathcal{B}_{i,n}(\epsilon) &:= \{\hat{\sigma}_{i,n} \in [\sigma_i, \sigma_i + \epsilon]\} \\ \mathcal{E}_{i,n}(\epsilon) &:= \mathcal{B}_{i,n}(\epsilon) \cap \mathcal{A}_{i,n}(\epsilon),\end{aligned}$$

where  $n \in \mathbb{N}$ , and

$$\epsilon_{i,l}(\epsilon) = \frac{\epsilon \alpha_i^2}{1 + \epsilon \alpha_i}, \quad \epsilon_{i,u}(\epsilon) = \frac{\epsilon \alpha_i^2 (\sigma_i + 1)}{\sigma_i - \epsilon \alpha_i (\sigma_i + 1)}. \quad (4.17)$$

Note that  $\bar{\alpha}_i(n) = \hat{\alpha}_{i,n}$  holds on  $\mathcal{A}_{i,n}(\epsilon)$  for any  $n \geq \alpha_i + 1$ . In Theorem 4.2, we set  $\varepsilon_i$  to satisfy  $\hat{\mu}_i \in [\mu_i - \delta_i, \mu_i + \delta_i]$  on  $\mathcal{E}_i(\epsilon)$  for any  $\epsilon \leq \varepsilon_i$ . Then, (GO) is decomposed by

$$\begin{aligned}(\text{GO}) &= \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] \\ &\quad + \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}^c(\epsilon), \mathcal{M}_\epsilon(t)] \\ &\leq \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}(\epsilon), \mathcal{M}_\epsilon(t)] + \Delta_i \mathbb{1} [i(t) = i, \mathcal{E}_{i,N_i(t)}^c(\epsilon)].\end{aligned}$$

Then, two lemmas below conclude the proof of Lemma 4.5, whose proofs are postponed to Section 4.6.6.

**Lemma 4.7.** *Under TS and TS-T with  $k \in \mathbb{Z}$ , it holds that for any  $i \in [K]$*

$$\sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = i, \mathcal{M}_\epsilon(t), \mathcal{E}_{i,N_i(t)}(\epsilon)]] \leq \max(0, k) + 1 + \frac{1}{\alpha_i \epsilon} \mathbb{1}[k > 0] + \frac{\log T}{D_{i,k}(\epsilon)}.$$

where  $D_{i,k}(\epsilon) > 0$  is defined in (4.11).

**Lemma 4.8.** *Under TS and TS-T with  $k \in \mathbb{Z}$ , it holds that for any  $i \neq 1$*

$$\sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = i, \mathcal{E}_{i,N_i(t)}^c(\epsilon)]] \leq \mathcal{O}(\epsilon^{-2}).$$

Lemma 4.8 controls the regret induced when estimators of the played arm are not close to their true parameters, which is not difficult to analyze as in the usual analysis of TS. In fact, the proof of this lemma is straightforward since the upper bounds of  $\mathbb{P}[\mathcal{B}_i^c]$  and  $\mathbb{P}[\mathcal{B}_i, \mathcal{A}_i^c]$  can be easily derived based on the distributions of  $\hat{\sigma}_{i,n}$  and  $\hat{\alpha}_{i,n}$  in (4.5).  $\square$

#### 4.6.5 Proof of Lemma 4.6

*Proof.* (BO) is decomposed by

$$\begin{aligned}(\text{BO}) &= \sum_{i=2}^K \sum_{t=Kn_0+1}^T \Delta_i \mathbb{1} [i(t) = i, \mathcal{B}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t)] \\ &\quad + \Delta_i \mathbb{1} [i(t) = i, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)].\end{aligned}$$

Then, two lemmas below conclude the proof of Lemma 4.6, whose proofs are postponed to Section 4.6.7.

**Lemma 4.9.** *Under TS with  $k \in \mathbb{Z}_{\geq 2}$ ,*

$$\sum_{t=n_0K+1}^T \mathbb{E} \left[ \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)] \right] \leq \mathcal{O}(\epsilon^{-2}).$$

*and under TS-T with  $k \in \mathbb{Z}_{\geq 0}$ ,*

$$\sum_{t=n_0K+1}^T \mathbb{E} \left[ \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)] \right] \leq \mathcal{O}(\epsilon^{-m}),$$

*where  $m = \max(2, 3 - k)$ .*

**Lemma 4.10.** *Under TS and TS-T with  $k \in \mathbb{Z}_{\geq 0}$ ,*

$$\sum_{t=n_0K+1}^T \mathbb{E} \left[ \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)] \right] \leq \mathcal{O}(\epsilon^{-1}).$$

The key to Lemma 4.10 is to convert the term on  $\tilde{\mu}_1(t)$ ,  $\mathcal{M}_\epsilon(t)$ , to a term on  $\tilde{\alpha}_1(t)$ . Since  $\mu(\sigma, \alpha) = \infty$  holds for  $\alpha \leq 1$ ,  $\tilde{\mu}_1 = \mu(\tilde{\sigma}_1, \tilde{\alpha}_1)$  becomes infinity regardless of the value of  $\tilde{\sigma}_1$  if  $\tilde{\alpha}_1 \leq 1$  holds, which implies  $\mathbb{P}[\mathcal{M}_\epsilon^c(t), \tilde{\alpha}_1(t) \leq 1] = 0$ . Therefore, it is enough to consider the case where  $\tilde{\alpha}_1(t) > 1$  holds to prove Lemma 4.10. Although density functions of  $\tilde{\alpha}_1$  under TS and TS-T are different, conditional CDFs of  $\tilde{\sigma}_1$  given  $\alpha_1 = \tilde{\alpha}_1$  are the same, which is given in (4.7) as

$$\mathbb{P}[\tilde{\sigma}_1 \leq x | T(X_{1,N_1(t)}), \tilde{\alpha}_1 = \alpha_1] = \left( \frac{x}{\hat{\sigma}_{1,n}(N_1(t))} \right)^{\tilde{\alpha}_1 N_1(t)}.$$

Therefore, for sufficiently large  $N_1(t)$  and  $\tilde{\alpha}_1(t) > 1$ ,  $\tilde{\sigma}_1(t)$  will concentrate on  $\hat{\sigma}_{1,n}(N_1(t))$  with high probability, which is close to its true value  $\sigma_1$  under the condition  $\{\mathcal{B}_{1,N_1(t)}(\epsilon)\}$ . Thus,  $\tilde{\mu}_1 = \frac{\tilde{\sigma}_1 \tilde{\alpha}_1}{\tilde{\alpha}_1 - 1} \geq \frac{\sigma_1 \tilde{\alpha}_1}{\tilde{\alpha}_1 - 1} = \mu(\sigma_1, \tilde{\alpha}_1)$  holds with high probability, which implies

$$\mathbb{P}[\mathcal{B}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t) | T(X_{1,N_1(t)})] \lesssim \mathbb{P}[\mathcal{B}_{1,N_1(t)}(\epsilon), \tilde{\alpha}_1(t) \geq c | T(X_{1,N_1(t)})]$$

for some problem-dependent constants  $c > 1$ . Since  $\mathcal{B}_1$  is deterministic by  $T(X_{1,N_1(t)})$ , we have

$$\mathbb{P}[\mathcal{B}_{1,N_1(t)}(\epsilon), \tilde{\alpha}_1(t) \geq c | T(X_{1,N_1(t)})] = \mathbb{1}[\mathcal{B}_{1,N_1(t)}(\epsilon)] \mathbb{P}[\tilde{\alpha}_1(t) \geq c | T(X_{1,N_1(t)})],$$

which implies  $\tilde{\mu}_1(t)$  is mainly determined by the value of  $\tilde{\alpha}_1(t)$  under the condition  $\{\mathcal{B}_{1,N_1(t)}(\epsilon)\}$  for both policies. In such cases, TS and TS-T behave like TS in the Pareto distribution with a known scale parameter, where  $\tilde{\mu}_1(t) := \mu(\sigma_1, \tilde{\alpha}_1(t))$  for  $t \in \mathbb{N}$ . Here, the Pareto distribution with the known scale parameter belongs to the one-dimensional exponential family, where Korda et al. [2013] showed the optimality of TS with the Jeffreys prior. Since the posterior of  $\alpha$  under the Jeffreys prior is given as the Erlang distribution with shape  $N_1(t) + 1$  in the one-parameter Pareto model, we can apply the results by Korda et al. [2013] to prove Lemma 4.10 by using some properties of the Erlang distribution such as (4.10).  $\square$

#### 4.6.6 Proofs of technical lemmas for Lemma 4.5

In this section, we provide the detailed proofs of Lemmas 4.7 and 4.8. Before beginning the proof, we state a relation between the Erlang distribution and the chi-squared distribution and a fundamental inequality for the chi-squared distribution.

**Fact 4.11.** When  $X \sim \text{Erlang}(n, \beta)$  with rate parameter  $\beta$ , then  $2\beta X$  follows the chi-squared distribution with  $2n$  degree of freedom, i.e.,  $2\beta X \sim \chi_{2n}^2$ .

**Lemma 4.12.** Let  $Z$  be a random variable following the chi-squared distribution with the degree of freedom  $2n$ . Then, for any  $x \in (0, 1)$

$$\mathbb{P}[Z \leq 2nx] \leq e^{-nh(x)},$$

where  $h(x) = (x - 1 - \log x) \geq 0$ .

*Proof.* Let  $X_i$  be random variables following the standard normal distribution so that  $Z = \sum_{i=1}^{2n} X_i^2$  holds. From the Cramér's theorem (given in Lemma 4.20), one can derive

$$\mathbb{P}[Z \leq 2nx] = \mathbb{P}\left[\frac{1}{2n} \sum_{i=1}^{2n} X_i^2 \leq x\right] \leq \exp\left\{\left(-2n \inf_{z \leq x} \Lambda^*(z)\right)\right\}.$$

From the definition of the moment-generating function, one can see that

$$\Lambda^*(z) = \sup_{\lambda \in \mathbb{R}} \lambda z - \log \mathbb{E}\left[e^{\lambda X_1^2}\right] = \sup_{\lambda \in \mathbb{R}} \lambda z + \frac{1}{2} \log(1 - 2\lambda) = \frac{1}{2}(z - 1 - \log z),$$

which concludes the proof.  $\square$

To avoid redundancy, we use a temporary notation  $\alpha_{i,n}$  when the same result holds for both  $\hat{\alpha}_{i,n}$  and  $\bar{\alpha}_{i,n}$ . When  $\alpha_{i,n}$  notation is used, one can replace it with either  $\hat{\alpha}_{i,n}$  or  $\bar{\alpha}_{i,n}$  depending on which policy we are considering. For example, it holds that

$$\alpha_{i,n} \leq 1 \Leftrightarrow \begin{cases} \hat{\alpha}_{i,n} \leq 1 & \text{under TS policy,} \\ \bar{\alpha}_{i,n} \leq 1 & \text{under TS-T policy.} \end{cases}$$

Similarly, we use the notation  $\theta_{i,n} := (\hat{\sigma}_{i,n}, \alpha_{i,n})$  when it can be replaced by both  $\hat{\theta}_{i,n} = (\hat{\sigma}_{i,n}, \hat{\alpha}_{i,n})$  and  $\bar{\theta}_{i,n} = (\hat{\sigma}_{i,n}, \bar{\alpha}_{i,n})$  for any arm  $a \in [K]$  and  $n \in \mathbb{N}$ . From here on out, we will use this notation throughout the remainder of this chapter.

*Proof of Lemma 4.7.* Fix a time index  $t$  and  $N_i(t) = n$ , and denote  $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid T(X_{i,N_i(t)})] = \mathbb{P}[\cdot \mid \theta_{i,n}]$ . To simplify notations, we drop the argument  $t$  of  $\tilde{\sigma}_i(t)$ ,  $\tilde{\alpha}_i(t)$ , and  $\tilde{\mu}_i(t)$ . From the sampling rule and the definition of  $\mathcal{M}_\epsilon$ , it holds for any  $i \neq 1$  that

$$\begin{aligned} \mathbb{E}[\mathbb{1}[i(t) = i, \tilde{\mu}^*(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t) = n]] \\ \leq \mathbb{E}_{\theta_{i,n}}[\mathbb{P}_t[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon), N_i(t) = n]]. \end{aligned}$$

Since  $\tilde{\sigma}_i \in (0, \hat{\sigma}_{i,n}]$  holds from its posterior distribution, if  $\tilde{\alpha}_i \geq \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}$  holds, then  $\tilde{\mu}_i = \frac{\tilde{\sigma}_i \tilde{\alpha}_i}{\tilde{\alpha}_i - 1} \leq \mu_1 - \epsilon$  holds. Recall that  $f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(\cdot)$  denotes a density function of  $\text{Erlang}\left(n - k, \frac{n}{\alpha_{i,n}}\right)$  with rate parameter  $\frac{n}{\alpha_{i,n}}$ , which is the marginalized posterior distribution of  $\tilde{\alpha}_i$  under TS and TS-T. From the CDF of  $\tilde{\sigma}$  in (4.7), if  $\hat{\sigma}_{i,n} < \mu_1 - \epsilon$ , then

$$\begin{aligned} \mathbb{P}_t[\tilde{\mu}_i \geq \mu_1 - \epsilon] &= \mathbb{P}_t[\tilde{\alpha}_i \leq 1] + \mathbb{P}_t\left[\tilde{\sigma}_i \geq \frac{\tilde{\alpha}_i - 1}{\tilde{\alpha}_i}(\mu_1 - \epsilon) \cap \tilde{\alpha}_i \in \left(1, \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}\right)\right] \\ &= \int_0^1 f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx + \int_1^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \mathbb{P}_t\left[\tilde{\sigma}_i \geq \frac{x - 1}{x}(\mu_1 - \epsilon)\right] dx \\ &= \int_0^1 f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx + \int_1^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left(1 - \left(\frac{x - 1}{\hat{\sigma}_{i,n} x}(\mu_1 - \epsilon)\right)^{nx}\right) dx \\ &= \int_0^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx - \int_1^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left(\frac{x - 1}{\hat{\sigma}_{i,n} x}(\mu_1 - \epsilon)\right)^{nx} dx. \end{aligned}$$

Since  $\frac{x}{x-y}$  is increasing with respect to  $y < x$  and  $\hat{\sigma}_{i,n} \leq \sigma_i + \epsilon$  holds on  $\mathcal{E}_{i,n}$ , we have on  $\mathcal{E}_{i,n}$  that

$$\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}} \leq \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - (\sigma_i + \epsilon)}.$$

Let

$$\eta_i(\epsilon) = \frac{\sigma_i(\Delta_i - \epsilon) - \epsilon\mu_i}{(\mu_i - \sigma_i)(\mu_1 - \sigma_i - 2\epsilon)} > 0$$

be a deterministic constant that depends only on the model and  $\epsilon$  satisfying

$$\begin{aligned} \alpha_i - \eta_i(\epsilon) &= \frac{\mu_i}{\mu_i - \sigma_i} - \frac{\sigma_i(\Delta_i - \epsilon) - \epsilon\mu_i}{(\mu_i - \sigma_i)(\mu_1 - \sigma_i - 2\epsilon)} \\ &= \frac{\mu_i\mu_1 - \sigma_i\mu_i - 2\epsilon\mu_i - \sigma_i(\mu_1 - \mu_i - \epsilon) + \epsilon\mu_i}{(\mu_i - \sigma_i)(\mu_1 - \sigma_i - 2\epsilon)} \\ &= \frac{\mu_1(\mu_i - \sigma_i) - \epsilon(\mu_i - \sigma_i)}{(\mu_i - \sigma_i)(\mu_1 - \sigma_i - 2\epsilon)} \\ &= \frac{\mu_1 - \epsilon}{\mu_1 - \sigma_i - 2\epsilon}. \end{aligned}$$

Since  $\eta_i(\epsilon) > 0$ , it holds that for any  $\epsilon \in (0, \varepsilon_i)$

$$\alpha_i - \eta_i(\epsilon) = \frac{\mu_1 - \epsilon}{\mu_1 - \sigma_i - 2\epsilon} \leq \frac{\mu_i}{\mu_i - \sigma_i} = \alpha_i. \quad (4.18)$$

Notice that  $\mu(\sigma, \alpha) = \frac{\sigma\alpha}{\alpha-1} = \sigma + \frac{\sigma}{\alpha-1}$  holds, which gives  $\frac{\mu}{\mu-\sigma} = \alpha$ . Therefore, the change of  $\mu$  to  $\mu'$  with fixed  $\sigma$ ,  $\frac{\mu'}{\mu'-\sigma}$ , implies how the value of the shape parameter  $\alpha'$  should be to satisfy  $\mu((\sigma, \alpha')) = \mu'$ . For example,  $\theta = (\sigma_i + \varepsilon_i, \alpha_i)$  satisfies  $\mu(\theta) \leq \mu_i + \frac{\delta_i}{2}$ . Thus, if  $\mu((\sigma_i + \varepsilon_i, \alpha)) = \mu_1 - \epsilon > \mu_i + \frac{\delta_i}{2}$ , then  $\alpha$  should be smaller than  $\alpha_i$ . Hence, we have

$$\begin{aligned} \mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] \mathbb{P}_t \left[ \tilde{\mu}_i \geq \mu_1 - \epsilon \right] &\leq \mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] \left( \int_0^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \right. \\ &\quad \left. - \int_1^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left( \frac{x-1}{\hat{\sigma}_{i,n}x} (\mu_1 - \epsilon) \right)^{nx} dx \right) \\ &\leq \mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] \int_0^{\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \hat{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \\ &\leq \mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] \int_0^{\alpha_i - \eta_i(\epsilon)} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \\ &= \mathbb{1}[\mathcal{E}_{i,n}(\epsilon)] \mathbb{P}_t[\tilde{\alpha}_i(t) \leq \alpha_i - \eta_i(\epsilon)]. \end{aligned}$$

Therefore, by taking expectations and using Fact 4.11, we have

$$\begin{aligned} \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] &\leq \mathbb{P}[\tilde{\alpha}_i \leq \alpha_i - \eta_i(\epsilon), \mathcal{E}_{i,n}(\epsilon)], \\ &= \mathbb{P} \left[ Z \leq \frac{2n}{\alpha_{i,n}} (\alpha - \eta_i(\epsilon)), \mathcal{E}_{i,n}(\epsilon) \right] \end{aligned} \quad (4.19)$$

where  $Z$  is a random variable following the chi-squared distribution with  $2(n-k)$  degrees of freedom, i.e.,  $Z \sim \chi_{2n-2k}^2$ .

## Under TS

Here, we first consider the case of TS where we replace  $\alpha_{i,n}$  with  $\hat{\alpha}_{i,n}$ .

Since  $\hat{\alpha}_{i,n} \in [\alpha_i - \epsilon_{i,l}, \alpha_i + \epsilon_{i,u}]$  holds on  $\mathcal{E}_{i,n}(\epsilon)$ , we have

$$\frac{1}{\alpha_i} - \epsilon \left(1 + \frac{1}{\sigma_i}\right) = \frac{1}{\alpha_i + \epsilon_{i,u}} \leq \frac{1}{\hat{\alpha}_{i,n}} \leq \frac{1}{\alpha_i - \epsilon_{i,l}} = \frac{1}{\alpha_i} + \epsilon \quad (4.20)$$

by the definition of  $\epsilon_{i,l}(\epsilon)$  and  $\epsilon_{i,u}(\epsilon)$  in (4.17). By replacing  $\alpha_{i,n}$  with  $\hat{\alpha}_{i,n}$  in (4.19) and applying (4.20), we have

$$\begin{aligned} \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] &\leq \mathbb{P}\left[Z \leq \frac{2n}{\hat{\alpha}_{i,n}} (\alpha_i - \eta_i(\epsilon)), \mathcal{E}_{i,n}(\epsilon)\right] \\ &\leq \mathbb{P}\left[Z \leq 2n \left(\frac{1}{\alpha_i} + \epsilon\right) (\alpha_i - \eta_i(\epsilon))\right] \\ &= \mathbb{P}\left[Z \leq 2(n-k) \frac{n}{n-k} \left(\frac{1}{\alpha_i} + \epsilon\right) (\alpha_i - \eta_i(\epsilon))\right]. \end{aligned}$$

**Priors with nonpositive  $k$**  Let us first consider the case  $k \in \mathbb{Z}_{\leq 0}$ , where we have  $\frac{n}{n-k} \leq 1$ . It holds that

$$\begin{aligned} \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] &\leq \mathbb{P}\left[Z \leq 2(n-k) \frac{n}{n-k} \left(\frac{1}{\alpha_i} + \epsilon\right) (\alpha_i - \eta_i(\epsilon))\right] \\ &\leq \mathbb{P}\left[Z \leq 2(n-k) \left(\frac{1}{\alpha_i} + \epsilon\right) (\alpha_i - \eta_i(\epsilon))\right]. \end{aligned}$$

Note that the definition of  $\varepsilon_i$  in Theorem 4.2 is set to satisfy  $\left(\frac{1}{\alpha_i} + \epsilon\right) (\alpha_i - \eta_i(\epsilon)) < 1$  for any  $\epsilon \leq \varepsilon_i$ . Thus, we can apply Lemma 4.12, which shows

$$\mathbb{P}\left[Z \leq 2(n-k) \left(1 - \frac{\eta_i(\epsilon)}{\alpha_i} + \epsilon(\alpha_i - \eta_i(\epsilon))\right)\right] \leq e^{-(n-k)D_{i,k}(\epsilon)}, \quad (4.21)$$

where

$$\begin{aligned} D_{i,k}(\epsilon) &:= -\log \left(1 - \frac{\eta_i(\epsilon)}{\alpha_i} + (\max(0, k) + 1)\epsilon(\alpha_i - \eta_i(\epsilon))\right) \\ &\quad - \frac{\eta_i(\epsilon)}{\alpha_i} + (\max(0, k) + 1)\epsilon(\alpha_i - \eta_i(\epsilon)) \end{aligned} \quad (4.22)$$

is a finite positive constant that only depends on the model parameters,  $\epsilon$ , and prior parameter  $k$ .

For arbitrary  $n_i > 0$ , applying (4.21) to (4.19) gives

$$\begin{aligned} \sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = a, \tilde{\mu}_1(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon)]] \\ &\leq \sum_{t=n_0K+1}^T \mathbb{P}[i(t) = a, \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] \\ &\leq n_i + \sum_{t=n_0K+1}^T \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t) \geq n_i] \\ &\leq n_i + \sum_{t=n_0K+1}^T e^{-(n_i-k)D_{i,k}(\epsilon)} \\ &\leq n_i + \sum_{t=n_0K+1}^T e^{-n_i D_{i,k}(\epsilon)} = n_i + T e^{-n_i D_{i,k}(\epsilon)}. \end{aligned}$$

Letting  $n_i = \frac{\log T}{D_{i,k}(\epsilon)}$  concludes the cases of priors with  $k \in \mathbb{Z}_{\leq 0}$ .

**Priors with positive  $k$**  Next, consider the case  $k \in \mathbb{Z}_{>0}$ . Recall that we first play every arm  $k+1$  times if  $k > 0$ , which implies that  $n - k > 0$ . For  $n \geq \frac{1}{\alpha\epsilon} + k + 1$ , it holds that

$$\frac{n}{n-k} \left( \frac{1}{\alpha} + \epsilon \right) \leq \frac{1}{\alpha} + (k+1)\epsilon. \quad (4.23)$$

By applying (4.23) to (4.19), we have for  $n \geq \frac{1}{\alpha\epsilon} + k + 1$ ,

$$\begin{aligned} \mathbb{P}[\tilde{\alpha}_i \leq \alpha_i - \eta_i(\epsilon), \mathcal{E}_{i,N_i(t)}(\epsilon)] \\ \leq \mathbb{P} \left[ Z \leq 2(n-k) \left( 1 - \frac{\eta_i(\epsilon)}{\alpha_i} + (k+1)\epsilon(\alpha_i - \eta_i(\epsilon)) \right) \right]. \end{aligned}$$

Similarly, by applying Lemma 4.12, one can see that for  $n \geq \frac{1}{\alpha_i\epsilon} + k + 1$

$$\mathbb{P}[\tilde{\alpha}_i \leq \alpha_i - \eta_i(\epsilon), \mathcal{E}_{i,N_i(t)}(\epsilon)] \leq e^{-(n-k)D_{i,k}(\epsilon)}, \quad (4.24)$$

where  $D_{i,k}(\epsilon)$  is defined in (4.22).

When  $k \in \mathbb{Z}_{>0}$ , let  $n_i \geq \frac{1}{\alpha_i\epsilon} + k + 1$  be arbitrary. By applying (4.24) to (4.19) again, we have

$$\begin{aligned} \sum_{t=n_0K+1}^T \mathbb{E}[\mathbb{1}[i(t) = a, \tilde{\mu}_1(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon)]] \\ \leq \sum_{t=n_0K+1}^T \mathbb{P}[i(t) = a, \tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] \\ \leq n_i + \sum_{t=n_0K+1}^T \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,N_i(t)}(\epsilon), N_i(t) \geq n_i] \\ \leq n_i + \sum_{t=n_0K+1}^T e^{-(n_i-k)D_{i,k}(\epsilon)} = n_i + Te^{-(n_i-k)D_{i,k}(\epsilon)}. \end{aligned}$$

Letting  $n_i = k + 1 + \frac{1}{\alpha_i\epsilon} + \frac{\log T}{D_{i,k}(\epsilon)}$  concludes the cases of priors with  $k > 0$ .

### Under TS-T

Here, we consider the case of TS-T where we replace  $\alpha_{i,n}$  with  $\bar{\alpha}_{i,n} = \min(\hat{\alpha}_{i,n}, n)$ . From the definition of  $\bar{\alpha}_{i,n}$ , it holds that for  $\epsilon \leq \varepsilon_i$

$$\forall n \geq \alpha_i + 1 : \mathbb{1}[\bar{\alpha}_{i,n} = \hat{\alpha}_{i,n}, \mathcal{A}_{i,n}(\epsilon)] = 1.$$

Therefore, for  $n \geq \alpha_i + 1$ , the analysis on TS can be applied to TS-T directly.

Let us consider the case where  $\bar{\alpha}_{i,n} = n < \alpha_i + 1$  holds under the condition  $\mathcal{A}_{i,n}(\epsilon)$ . By replacing  $\alpha_{i,n}$  with  $n$  in (4.19) and following the same steps as in (4.19) and (4.21), we have for any  $k \in \mathbb{Z}$ ,

$$\begin{aligned} \mathbb{P}[\tilde{\mu}_i(t) \geq \mu_1 - \epsilon, \mathcal{E}_{i,n}(\epsilon)] &\leq \mathbb{P} \left[ Z \leq \frac{2n}{n} (\alpha_i - \eta_i(\epsilon)), \mathcal{E}_{i,n}(\epsilon) \right] \\ &\leq \mathbb{P} \left[ Z \leq 2(n-k) \frac{1}{n-k} \left( \frac{1}{\alpha_i} + \epsilon \right) (\alpha_i - \eta_i(\epsilon)), \mathcal{E}_{i,n}(\epsilon) \right] \\ &\leq \mathbb{P} \left[ Z \leq 2(n-k) \left( \frac{1}{\alpha_i} + \epsilon \right) (\alpha_i - \eta_i(\epsilon)), \mathcal{E}_{i,n}(\epsilon) \right] \\ &\leq e^{-(n-k)D_{i,k}(\epsilon)}, \end{aligned}$$

where  $D_{i,k}(\epsilon)$  defined in (4.22). Therefore, the same result as that of TS can be obtained for TS-T.

### Meaning of problem dependent constant D

Let us rewrite the definition of  $D_{i,k}(\epsilon)$  as

$$\begin{aligned} D_{i,k}(\epsilon) &= -\log \left( 1 - \frac{\eta_i(\epsilon)}{\alpha_i} + (\max(0, k) + 1)\epsilon(\alpha_i - \eta_i(\epsilon)) \right. \\ &\quad \left. - \frac{\eta_i(\epsilon)}{\alpha_i} + (\max(0, k) + 1)\epsilon(\alpha_i - \eta_i(\epsilon)) \right) \\ &= -\log \left( \frac{(\alpha_i - \eta_i(\epsilon))(1 + (\max(0, k) + 1)\alpha_i\epsilon)}{\alpha_i} \right. \\ &\quad \left. + \frac{(\alpha_i - \eta_i(\epsilon))(1 + (\max(0, k) + 1)\alpha_i\epsilon)}{\alpha_i} - 1 \right) \\ &= \log \left( \alpha_i \frac{1}{\alpha_i - \eta_i(\epsilon)} \frac{1}{1 + (\max(0, k) + 1)\alpha_i\epsilon} \right. \\ &\quad \left. + \frac{(\alpha_i - \eta_i(\epsilon))(1 + (\max(0, k) + 1)\alpha_i\epsilon)}{\alpha_i} - 1 \right). \end{aligned}$$

By injecting the closed form of  $\alpha_i - \eta_i(\epsilon)$  given in (4.18), we have for  $b_{i,k}(\epsilon) = \frac{1}{1 + (\max(0, k) + 1)\alpha_i\epsilon}$  that

$$D_{i,k}(\epsilon) = \log \left( \alpha_i b_{i,k}(\epsilon) \frac{\mu_1 - \epsilon - (\sigma_i + \epsilon)}{\mu_1 - \epsilon} \right) + \frac{1}{\alpha_i b_{i,k}(\epsilon)} \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - (\sigma_i + \epsilon)} - 1.$$

From Lemma 4.1, one can observe that

$$D_{i,k}(\epsilon) = \inf_{\theta: \mu(\theta) > \mu_1 - \epsilon} \text{KL}(\text{Pa}(\sigma_i + \epsilon, \alpha_i b_{i,k}(\epsilon)), \text{Pa}(\theta)).$$

Therefore,  $D_{i,k}(\epsilon)$  illustrates the minimum discrimination information for the hypothesis that the arm  $i$  is optimal, where the optimal arm has the mean  $\mu_1 - \epsilon$ , and the scale and shape of the arm  $i$  are  $\sigma_i + \epsilon$  and  $\alpha_i b_{i,k}(\epsilon)$ .  $\square$

Before beginning the proof of Lemma 4.8, we introduce two lemmas on the event  $\mathcal{B}$  and  $\mathcal{A}$ , whose proofs are given after the proof of Lemma 4.8.

**Lemma 4.13.** *For any  $i \in [K]$ , it holds that for all  $\epsilon > 0$ ,  $t > 0$ , and  $n \in \mathbb{N}$*

$$\mathbb{P} \left[ \mathcal{B}_{i, N_i(t)}^c(\epsilon), N_i(t) = n \right] \leq \exp \left( -\frac{\alpha_i \epsilon}{\sigma_i + \epsilon} n \right).$$

**Lemma 4.14.** *For any  $i \in [K]$ , it holds that for all  $\epsilon \in \left( 0, \frac{\sigma_i}{\alpha_i(\sigma_i + 1)} \right)$  and  $t > 0$ , and  $n \geq n_0$*

$$\mathbb{P} \left[ \mathcal{A}_{i, N_i(t)}^c(\epsilon), \mathcal{B}_{i, N_i(t)}(\epsilon), N_i(t) = n \right] \leq 2 \exp \left( -\frac{\alpha_i^2 \epsilon^2}{4} n \right),$$

*Proof of Lemma 4.8.* From the Lemmas 4.13 and 4.14, one can see that for  $n \geq n_0$ ,

$$\begin{aligned} \mathbb{P} \left[ \mathcal{E}_{i, N_i(t)}^c(\epsilon), N_i(t) = n \right] &= \mathbb{P} \left[ \mathcal{B}_{i, N_i(t)}^c(\epsilon), N_i(t) = n \right] \\ &\quad + \mathbb{P} \left[ \mathcal{A}_{i, N_i(t)}^c(\epsilon), \mathcal{B}_{i, N_i(t)}(\epsilon), N_i(t) = n \right] \\ &\leq \exp \left( -\frac{\alpha_i \epsilon}{\sigma_i + \epsilon} n \right) + 2 \exp \left( -\frac{\alpha_i^2 \epsilon^2}{4} n \right). \end{aligned}$$



Since  $\{i(t) = i, \mathcal{E}_{i,n}^c(\epsilon), N_i(t) = n\}$  occurs only once for any  $n \in \mathbb{N}$ , it holds that

$$\begin{aligned}
& \sum_{t=n_0K+1}^T \mathbb{E} \left[ \mathbb{1} \left[ i(t) = i, \mathcal{E}_{i,N_i(t)}^c(\epsilon) \right] \right] \\
&= \sum_{t=n_0K+1}^T \sum_{n=n_0}^T \mathbb{E} \left[ \mathbb{1} \left[ i(t) = i, \mathcal{E}_{i,N_i(t)}^c(\epsilon), N_i(t) = n \right] \right] \\
&\leq \sum_{n=n_0}^{\infty} \mathbb{E} \left[ \mathbb{1} \left[ \mathcal{E}_{i,N_i(t)}^c(\epsilon), N_i(t) = n \right] \right] \\
&= \sum_{n=n_0}^{\infty} \mathbb{P} \left[ \mathcal{B}_{i,N_i(t)}^c(\epsilon), N_i(t) = n \right] \\
&\quad + \mathbb{P} \left[ \mathcal{A}_{i,N_i(t)}^c(\epsilon) \cap \mathcal{B}_{i,N_i(t)}^c(\epsilon), N_i(t) = n \right] \\
&\leq \sum_{n=n_0}^{\infty} \exp \left( -\frac{\alpha_i \epsilon}{\sigma_i + \epsilon} n \right) + 2 \exp \left( -\frac{\alpha_i^2 \epsilon^2}{4} n \right).
\end{aligned}$$

Since  $\exp(-xn)$  with  $x > 0$  is a decreasing function with respect to  $n$ , it holds that

$$\sum_{n=2}^{\infty} \exp(-xn) \leq \int_1^{\infty} \exp(-xn) dn = \frac{\exp(-x)}{x},$$

which concludes the proof.  $\square$

#### Proof of Lemma 4.13

*Proof.* Since  $\hat{\sigma}_{i,n}(n) \sim \text{Pareto}(\sigma_i, n\alpha_i)$  holds for any  $n \in \mathbb{N}$  in (4.5), it holds that

$$\begin{aligned}
\mathbb{P} \left[ \mathcal{B}_{i,N_i(t)}^c, N_i(t) = n \right] &= \mathbb{P} \left[ \hat{\sigma}_{i,n}(N_i(t)) \geq \sigma_i + \epsilon, N_i(t) = n \right] \\
&= \left( \frac{\sigma_i}{\sigma_i + \epsilon} \right)^{n\alpha_i} \leq \exp \left( -\frac{\alpha_i \epsilon}{\sigma_i + \epsilon} n \right),
\end{aligned}$$

which concludes the proof.  $\square$

#### Proof of Lemma 4.14

Although one could derive an upper bound based on the sampling distribution of  $\alpha$  given in (4.5) similarly to the proof of Lemma 4.13, we utilize the relation between the Pareto distribution and the exponential distribution in Fact 4.15 since we found it is more convenient to control in our analysis.

**Fact 4.15.** When  $X \sim \text{Pareto}(\sigma, \alpha)$  with the scale parameter  $\sigma \in \mathbb{R}_+$  and rate parameter  $\alpha \in \mathbb{R}_+$ , then  $\log \left( \frac{X}{\sigma} \right)$  follows the exponential distribution with rate  $\alpha$ , i.e.,  $\log \left( \frac{X}{\sigma} \right) \sim \text{Exp}(\alpha)$ .

*Proof.* Fix a time index  $t$  and denote  $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot \mid T(X_{i,N_i(t)})]$  and  $N_i(t) = n$ . From the definition of MLE of  $\alpha_i$ ,

$$\begin{aligned}
& \mathbb{P}[\hat{\alpha}_{i,n} \leq \alpha_i - \epsilon_{i,l}(\epsilon), \mathcal{B}_{i,N_i(t)}(\epsilon), N_i(t) = n] \\
&\leq \mathbb{P} \left[ \frac{n}{\sum_{s=1}^n \log x_{i,s} - n \log x_i^{(1)}} \leq \alpha_i - \epsilon_{i,l}(\epsilon) \right].
\end{aligned}$$

Then, we have

$$\begin{aligned}
\mathbb{P}\left[\frac{n}{\sum_{s=1}^n \log x_{i,s} - n \log x_i^{(1)}} \leq \alpha_i - \epsilon_{i,l}(\epsilon)\right] &= \mathbb{P}\left[\frac{n}{\alpha_i - \epsilon_{i,l}(\epsilon)} \leq \sum_{s=1}^n \log \frac{x_{i,s}}{x_i^{(1)}}\right] \\
&= \mathbb{P}\left[\frac{n}{\alpha_i - \epsilon_{i,l}(\epsilon)} \leq n \log \frac{\sigma}{x_i^{(1)}} + \sum_{s=1}^n \log \frac{x_{i,s}}{\sigma}\right] \\
&\leq \mathbb{P}\left[\frac{n}{\alpha_i - \epsilon_{i,l}(\epsilon)} \leq \sum_{s=1}^n \log \frac{x_{i,s}}{\sigma_i}\right] \\
&\leq \mathbb{P}\left[\epsilon \leq \frac{1}{n} \sum_{s=1}^n \log \frac{x_{i,s}}{\sigma_i} - \frac{1}{\alpha_i}\right],
\end{aligned}$$

where the first equality holds from the definition of MLEs in (4.5), the first inequality holds since any sample generated from the Pareto distribution cannot be smaller than its scale parameter  $\sigma$ , and the last inequality holds from the definition of  $\epsilon_{i,l}(\epsilon)$  in (4.17).

Similarly, one can derive that

$$\begin{aligned}
\mathbb{P}[\hat{\alpha}_{i,n} \geq \alpha_i + \epsilon_{i,u}(\epsilon), \mathcal{B}_{i,N_i(t)}(\epsilon), N_i(t) = n] \\
\leq \mathbb{P}\left[\sum_{s=1}^n \log \frac{x_{i,s}}{\sigma_i} \leq \frac{n}{\alpha_i + \epsilon_{i,u}(\epsilon)} + n \log \frac{x_1^{(1)}}{\sigma} \cap \mathcal{B}_{i,n}\right] \\
\leq \mathbb{P}\left[\sum_{s=1}^n \log \frac{x_{i,s}}{\sigma} \leq \frac{n}{\alpha_i + \epsilon_{i,u}(\epsilon)} + n \log \frac{\sigma_i + \epsilon}{\sigma_i}\right] \\
\leq \mathbb{P}\left[\sum_{s=1}^n \log \frac{x_{i,s}}{\sigma_i} \leq \frac{n}{\alpha_i + \epsilon_{i,u}(\epsilon)} + \frac{n\epsilon}{\sigma_i}\right] \\
\leq \mathbb{P}\left[\frac{1}{n} \sum_{s=1}^n \log \frac{x_{i,s}}{\sigma_i} - \frac{1}{\alpha_i} \leq -\epsilon\right],
\end{aligned}$$

where the second inequality holds since  $x_i^{(1)} = \hat{\sigma}_{i,n} \leq \sigma_i + \epsilon$  holds on  $\mathcal{B}_{i,n}$ , the third inequality from  $\log(1+x) \leq x$  for  $x > -1$ , and the last inequality comes from the definition of  $\epsilon_{i,u}(\epsilon)$ . From Fact 4.15,  $y_{i,s} := \log\left(\frac{x_{i,s}}{\sigma_i}\right) \sim \text{Exp}(\alpha_i)$  and the last probability can be considered as a deviation of the sum of exponentially distributed random variables.

For the exponential distribution  $\text{Exp}(\alpha)$ , we say that Bernstein's condition with parameter  $b$  holds if

$$\mathbb{E}[M_k] \leq \frac{1}{2} k! \frac{1}{\alpha^2} b^{k-2} \quad \text{for } k = 3, 4, \dots,$$

where  $M_k$  implies the  $k$ -th central moment. For  $\text{Exp}(\alpha_i)$ , it holds that

$$\mathbb{E}[M_k] = \frac{!k}{\alpha_i^k} \leq \frac{k!}{2} \frac{1}{\alpha_i^2} \left(\frac{1}{\alpha_i}\right)^{k-2},$$

where  $!k$  is the subfactorial of  $k$  such that  $!k \leq \frac{k!}{e} + \frac{1}{2} \leq \frac{k!}{2}$  for  $k \geq 3$ . Hence, the exponential distribution with parameter  $\alpha_i$  satisfies Bernstein's condition with parameter  $\frac{1}{\alpha_i}$ , so that it is subexponential with parameters  $\left(\frac{2}{\alpha_i^2}, \frac{2}{\alpha_i}\right)$ . Therefore, by applying Bernstein's inequality (given in Lemma 4.21), we have

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{s=1}^n y_{i,s} - \frac{1}{\alpha_i}\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{n}{4} \min\{\alpha_i^2 \epsilon^2, \alpha_i \epsilon\}\right).$$

Note that it holds for  $\epsilon < \frac{\sigma_i}{\alpha_i(\sigma_i+1)}$  that

$$\begin{aligned}\mathbb{P}[\hat{\alpha}_{i,n} \leq \alpha_i - \epsilon_{i,l}(\epsilon) \cap \mathcal{B}_{i,n}] &\leq \mathbb{P}\left(\frac{1}{n} \sum_{s=1}^n y_{i,s} - \frac{1}{\alpha_i} \geq \epsilon\right) \\ \mathbb{P}[\hat{\alpha}_{i,n} \geq \alpha_i + \epsilon_{i,u}(\epsilon) \cap \mathcal{B}_{i,n}] &\leq \mathbb{P}\left(\frac{1}{n} \sum_{s=1}^n y_{i,s} - \frac{1}{\alpha_i} \leq -\epsilon\right),\end{aligned}$$

for  $\epsilon_{i,l}(\epsilon) = \frac{\epsilon \alpha_i^2}{1 + \epsilon \alpha_i}$  and  $\epsilon_{i,u}(\epsilon) = \frac{\epsilon \alpha_i^2(\sigma_i+1)}{\sigma_i - \epsilon \alpha_i(\sigma_i+1)}$ . Hence, we obtain

$$\begin{aligned}\mathbb{P}[\mathcal{A}_{i,N_i(t)}^c(\epsilon), \mathcal{B}_{i,N_i(t)}(\epsilon), N_i(t) = n] \\ = \mathbb{P}[\hat{\alpha}_{i,n}(n) \leq \alpha_i - \epsilon_{i,l}(\epsilon), \mathcal{B}_{i,N_i(t)}, N_i(t) = n] \\ + \mathbb{P}[\hat{\alpha}_{i,n}(n) \geq \alpha_i + \epsilon_{i,u}(\epsilon), \mathcal{B}_{i,N_i(t)}, N_i(t) = n] \\ \leq 2 \exp\left(-\frac{\alpha_i^2 \epsilon^2}{4} n\right),\end{aligned}$$

for  $\epsilon < \frac{1}{\alpha_i}$  with  $\alpha_i > 1$ . □

#### 4.6.7 Proofs of technical lemmas for Lemma 4.6

This section presents the detailed proofs of Lemmas 4.9 and 4.10.

Before beginning the proof of Lemma 4.9, we provide a lemma based on  $\theta_{1,n}$  notation. We denote the probability of sample from the posterior distribution after  $n$  times playing is smaller than  $\mu_1 - x$  by

$$p_n(x|\theta_{1,n}) := \mathbb{P}[\tilde{\mu}_1 \leq \mu_1 - x | \hat{\sigma}_{1,n}(n), \alpha_{1,n}]. \quad (4.25)$$

Let  $K(\epsilon) = (\sigma_1 + \epsilon, \mu_1 - \epsilon)$  be an open interval on  $\mathbb{R}$ . The Lemma 4.16 below shows the upper bound of  $p_n$  conditioned on  $\theta_{1,n}$ .

**Lemma 4.16.** *Given  $\epsilon > 0$ , define a positive problem-dependent constant  $\rho = \rho_{\theta_1}(\epsilon) := \frac{\sigma_1 \epsilon}{2(\mu_1 - \epsilon/2 - \sigma_1)(\mu_1 - \sigma_1)}$ . If  $n \geq n_0 = \max(2, k+1)$ , then for  $k \in \mathbb{Z}_{\geq 0}$*

$$p_n(\epsilon|\theta_{1,n}) \leq \begin{cases} e^{-n}, & \text{if } \hat{\sigma}_{1,n}(n) \geq \mu_1 - \epsilon, \\ h(\mu_1, \epsilon, n), & \text{if } \hat{\sigma}_{1,n}(n) \in K(\epsilon), \alpha_{1,n} \leq \alpha_1 + \rho, \\ C_1(\mu_1, \epsilon, n) G_k(1/\alpha_{1,n}; n) + 1 - G_k(1/\alpha_{1,n}; n) & \text{if } \hat{\sigma}_{1,n}(n) \in K(\epsilon), \alpha_{1,n} \geq \alpha_1 + \rho, \end{cases}$$

where

$$\begin{aligned}h(\mu_1, \epsilon, n) &:= e^{-n \frac{3\epsilon}{4\mu_1}} \left(1 - \frac{1}{2} e^{-nc_{\mu_1}(\epsilon)}\right) + \frac{1}{2} e^{-nc_{\mu_1}(\epsilon)} \\ C_1(\mu_1, \epsilon, n) &:= \left(\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon/2}\right)^n \leq e^{-n \frac{\epsilon}{2\mu_1 - \epsilon}} < 1 \\ G_k(x; n) &:= F_{n-k, nx}^{\text{Er}}(\alpha_1 + \rho)\end{aligned}$$

for  $F^{\text{Er}}$  defined in (4.9). Here,  $c_{\mu_1}(\epsilon) = \zeta - \log(1 + \zeta) = \mathcal{O}(\epsilon^{-2})$  and  $\zeta = \frac{\epsilon}{4\mu_1 - 2\epsilon} \in (0, 1)$  are deterministic constants of  $\mu_1$  and  $\epsilon$ .

Notice that  $\mu((\sigma_1, \alpha_1 + \rho)) = \mu_1 - \epsilon/2$  holds and there exists a problem-dependent constant  $C_2(\mu_1, \epsilon, k) < 1$  such that for any  $n \geq n_0 = \max(2, k+1)$  and  $\epsilon > 0$

$$h(\mu_1, \epsilon, n) \leq 1 - C_2(\mu_1, \epsilon, k). \quad (4.26)$$

Then, the following fact shows the relation between the Erlang distribution and the inverse gamma distribution.

**Fact 4.17.** When  $X \sim \text{Erlang}(n, \beta)$  with rate parameter  $\beta$ , then  $\frac{1}{X}$  follows the inverse gamma distribution with shape  $n \in \mathbb{N}$  and scale  $\beta \in \mathbb{R}_+$ , i.e.,  $\frac{1}{X} \sim \text{InvG}(n, \beta)$ .

*Proof of Lemma 4.9.* Let us consider the following decomposition that holds under both TS and TS-T:

$$\begin{aligned} & \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)] \\ &= \sum_{n=n_0}^T \sum_{t=Kn_0+1}^T \mathbb{1}\left[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t), N_1(t) = n\right] \\ &= \sum_{n=n_0}^T \sum_{m=1}^T \mathbb{1}\left[m \leq \sum_{t=Kn_0+1}^T \mathbb{1}\left[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), \mathcal{B}_{1,N_1(t)}^c(\epsilon), N_1(t) = n\right]\right]. \end{aligned}$$

Notice that

$$m \leq \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t), N_1(t) = n]$$

implies that  $\tilde{\mu}_1(t) \leq \mu_1 - \epsilon$  occurred  $m$  times in a row on  $\{t : \mathcal{B}_{1,n}^c(\epsilon), \mathcal{M}_\epsilon^c(t), N_1(t) = n\}$ . Thus, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)]\right] &\leq \mathbb{E}\left[\sum_{n=n_0}^T \sum_{m=1}^T \mathbb{1}[\mathcal{B}_{1,n}^c(\epsilon)] p_n(\epsilon|\theta_{1,n})^m\right] \\ &\leq \sum_{n=n_0}^T \mathbb{E}\left[\mathbb{1}[\mathcal{B}_{1,n}^c(\epsilon)] \frac{p_n(\epsilon|\theta_{1,n})}{1 - p_n(\epsilon|\theta_{1,n})}\right] \end{aligned} \quad (4.27)$$

for  $p_n$  defined in (4.25).

**Under TS with  $k \in \mathbb{Z}_{\geq 2}$**

By applying Lemma 4.16 and (4.26) under TS with  $k \in \mathbb{Z}_{\geq 0}$ , we can decompose the expectation in (4.27) as

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}[\mathcal{B}_{1,n}^c(\epsilon)] \frac{p_n(\epsilon|\hat{\theta}_{1,n})}{1 - p_n(\epsilon|\hat{\theta}_{1,n})}\right] &\leq \mathbb{P}[\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon] \frac{e^{-n}}{1 - e^{-n}} \\ &\quad + \mathbb{P}[\hat{\sigma}_{1,n} \in K(\epsilon), \hat{\alpha}_{1,n} < \alpha_1 + \rho] \frac{h(\mu_1, \epsilon, n)}{C_2(\mu_1, \epsilon, k)} \\ &\quad + \underbrace{\mathbb{E}_{\hat{\theta}_{1,n}}\left[\frac{\mathbb{1}[\hat{\sigma}_{1,n} \in K(\epsilon), \hat{\alpha}_{1,n} > \alpha_1 + \rho]}{G_k(1/\hat{\alpha}_{1,n}; n)(1 - C_{1,n})}\right]}_{(\star_P)}, \end{aligned} \quad (4.28)$$

where we denoted  $C_{1,n} = C_1(\mu_1, \epsilon, n)$ . For simplicity, let us define  $z := \frac{1}{\hat{\alpha}_{1,n}}$  where  $z \sim \text{Erlang}(n-1, n\alpha_1)$  holds from Fact (4.17) since  $\hat{\alpha}_{1,n} \sim \text{InvG}(n-1, n\alpha_1)$  in (4.5). From the independence of  $\hat{\sigma}$  and  $\hat{\alpha}$  and distributions of  $z$  and  $\hat{\alpha}$  in (4.8) and (4.5), respectively, we have

$$\begin{aligned} (\star_P) &= \int_0^{\frac{1}{\alpha_1+\rho}} z^{n-2} e^{-n\alpha_1 z} \frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)} \int_{\hat{\sigma}_{1,n} \in K(\epsilon)} \frac{f_{\sigma_1, n\alpha_1}^{\text{Pa}}(\hat{\sigma}_{1,n})}{G_k(z; n)(1 - C_{1,n})} d\hat{\sigma}_{1,n} dz \\ &= \mathbb{P}[\hat{\sigma}_{1,n} \in K(\epsilon)] \int_0^{\frac{1}{\alpha_1+\rho}} \frac{z^{n-2} e^{-n\alpha_1 z}}{G_k(z; n)(1 - C_{1,n})} \frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)} dz. \end{aligned}$$

By substituting the CDF in (4.9), we obtain

$$\begin{aligned}
G_k(z; n) &= F_{n-k, n}^{\text{Er}}(\alpha_1 + \rho) \\
&= \frac{1}{\Gamma(n-k)} \int_0^{n(\alpha_1 + \rho)z} e^{-t} t^{n-k-1} dt \\
&\geq \frac{e^{-n(\alpha_1 + \rho)z}}{\Gamma(n-k)} \int_0^{n(\alpha_1 + \rho)z} t^{n-k-1} dt \\
&= \frac{e^{-n(\alpha_1 + \rho)z}}{\Gamma(n-k+1)} (n(\alpha_1 + \rho)z)^{n-k}.
\end{aligned} \tag{4.29}$$

Therefore,

$$\begin{aligned}
&\frac{(\star_P)}{\mathbb{P}[\hat{\sigma} \in K(\epsilon)]} \\
&\leq \int_0^{\frac{1}{\alpha_1 + \rho}} \frac{\Gamma(n-k+1)}{(n(\alpha_1 + \rho)z)^{n-k}(1-C_{1,n})} e^{n(\alpha_1 + \rho)z} \frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)} z^{n-2} e^{-n\alpha_1 z} dz \\
&= \frac{\Gamma(n-k+1)}{\Gamma(n-1)(1-C_{1,n})} (\alpha_1 + \rho)^{k-1} \left( \frac{\alpha_1}{\alpha_1 + \rho} \right)^{n-1} n^{k-1} \int_0^{\frac{1}{\alpha_1 + \rho}} z^{k-2} e^{n\rho z} dz \\
&\leq \frac{\Gamma(n-k+1)}{\Gamma(n-1)(1-C_{1,n})} (\alpha_1 + \rho)^{k-1} e^{-\frac{\rho}{\alpha_1 + \rho}(n-1)} \frac{n^{k-1}}{(n\rho)^{k-2}} \int_0^{\frac{1}{\alpha_1 + \rho}} (n\rho z)^{k-2} e^{n\rho z} dz.
\end{aligned} \tag{4.30}$$

By letting  $n\rho z = y$ , we can bound the integral in (4.30) as

$$\begin{aligned}
\frac{n^{k-1}}{(n\rho)^{k-2}} \int_0^{\frac{1}{\alpha_1 + \rho}} (n\rho z)^{k-2} e^{n\rho z} dz &= \rho^{-(k-1)} \int_0^{\frac{n\rho}{\alpha_1 + \rho}} y^{k-2} e^y dy \\
&\leq \rho^{-(k-1)} e^{\frac{n\rho}{\alpha_1 + \rho}} \int_0^{\frac{n\rho}{\alpha_1 + \rho}} y^{k-2} dy
\end{aligned} \tag{4.31}$$

$$= \frac{e^{\frac{n\rho}{\alpha_1 + \rho}}}{k-1} \left( \frac{n}{\alpha_1 + \rho} \right)^{k-1}, \quad \text{if } k \in \mathbb{Z}_{\geq 2} \tag{4.32}$$

where (4.32) holds only for  $k \in \mathbb{Z}_{\geq 2}$  since the integral in (4.31) diverges for  $k \in \mathbb{Z}_{\leq 1}$ .

By applying (4.32) to (4.30), we obtain for  $k \in \mathbb{Z}_{\geq 2}$

$$\begin{aligned}
(\star_P) &\leq \mathbb{P}[\hat{\sigma} \in K(\epsilon)] \frac{e^{\frac{\rho}{\alpha_1 + \rho}}}{1-C_{1,n}} \frac{n^{k-1}}{k-1} \frac{\Gamma(n-k+1)}{\Gamma(n-1)} \\
&\leq \frac{e^{1-\frac{\epsilon\alpha_1}{\sigma+\epsilon}n}}{1-C_{1,n}} \frac{\Gamma(n-k+1)}{\Gamma(n-1)} \frac{n^{k-1}}{k-1} = \mathcal{O}(ne^{-n\epsilon}),
\end{aligned} \tag{4.33}$$

where (4.33) can be obtained by Lemma 4.13 and  $\frac{\rho}{\alpha_1 + \rho} < 1$ . By combining (4.33) with (4.28) and (4.27), we obtain for  $k \in \mathbb{Z}_{\geq 2}$

$$\begin{aligned}
&\mathbb{E} \left[ \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1, N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)] \right] \\
&\leq \sum_{n=n_0}^T \left( \frac{e^{-n}}{1-e^{-n}} + \frac{e^{-n\frac{3\epsilon}{4\mu_1}} + \frac{1}{2}e^{-nc\mu_1(\epsilon)}}{C_2(\mu_1, \epsilon, k)} + (\star_P) \right) \\
&\leq \sum_{n=n_0}^T \mathcal{O}(e^{-n}) + \mathcal{O}(e^{-n\epsilon}) + \mathcal{O}(e^{-n\epsilon^2}) + \mathcal{O}(ne^{-n\epsilon}) \\
&= \mathcal{O}(1) + \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-2}) + \mathcal{O}(\epsilon^{-2}),
\end{aligned}$$

which concludes the proof under TS with  $k \in \mathbb{Z}_{\geq 2}$ .

### Under TS-T with nonnegative $k$

Under TS-T, we have the following inequality instead of (4.28):

$$\begin{aligned} \mathbb{E} \left[ \mathbb{1} [\mathcal{B}_{1,n}^c(\epsilon)] \frac{p_n(\epsilon|\bar{\theta}_{1,n})}{1 - p_n(\epsilon|\bar{\theta}_{1,n})} \right] &\leq \mathbb{P}[\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon] \frac{e^{-n}}{1 - e^{-n}} \\ &\quad + \mathbb{P}[\hat{\sigma}_{1,n} \in K(\epsilon), \bar{\alpha}_{1,n} < \alpha_1 + \rho] \frac{h(\mu_1, \epsilon, n)}{C_2(\mu_1, \epsilon, k)} \\ &\quad + \underbrace{\mathbb{E}_{\bar{\theta}_{1,n}} \left[ \frac{\mathbb{1}[\hat{\sigma}_{1,n} \in K(\epsilon), \bar{\alpha}_{1,n} \in (\alpha_1 + \rho, n)]}{G_k(1/\bar{\alpha}_{1,n}; n)(1 - C_{1,n})} \right]}_{(\star'_P)}. \end{aligned} \quad (4.34)$$

From  $\mathbb{1}[\bar{\alpha}_{1,n}(n) < n] = \mathbb{1}[\bar{\alpha}_{1,n}(n) = \hat{\alpha}_{1,n}(n)]$ , it holds that

$$\begin{aligned} (\star'_P) &= \mathbb{E}_{\hat{\theta}_{1,n}} \left[ \frac{\mathbb{1}[\hat{\sigma}_{1,n} \in K(\epsilon), \hat{\alpha}_{1,n} \in (\alpha_1 + \rho, n)]}{G_k(1/\hat{\alpha}_{1,n}; n)(1 - C_{1,n})} \right] \\ &\quad + \mathbb{E}_{\bar{\theta}_{1,n}} \left[ \frac{\mathbb{1}[\hat{\sigma}_{1,n} \in K(\epsilon), \bar{\alpha}_{1,n} = n]}{G_k(1/\bar{\alpha}_{1,n}; n)(1 - C_{1,n})} \right]. \end{aligned}$$

Since  $\mathbb{1}[\bar{\alpha}_{1,n}(n) = n] = \mathbb{1}[\hat{\alpha}_{1,n}(n) \geq n]$  holds and  $\hat{\sigma}$  and  $\hat{\alpha}$  are independent, we have for  $z = \frac{1}{\hat{\alpha}_{1,n}} \sim \text{Erlang}(n-1, n\alpha_1)$

$$\begin{aligned} \frac{(\star'_P)}{\mathbb{P}[\hat{\sigma}_{1,n} \in K(\epsilon)]} &\leq \underbrace{\int_{\frac{1}{n}}^{\frac{1}{\alpha_1 + \rho}} \frac{z^{n-2} e^{-n\alpha_1 z}}{G_k(z; n)(1 - C_{1,n})} \frac{(n\alpha_1)^{n-1}}{\Gamma(n-1)} dz}_{(\dagger_P)} \\ &\quad + \underbrace{\frac{1}{G_k(1/n; n)(1 - C_{1,n})} \mathbb{P} \left[ \frac{1}{\hat{\alpha}_{1,n}} \leq \frac{1}{n} \right]}_{(\diamond_P)}, \end{aligned} \quad (4.35)$$

where the same techniques on  $(\star_P)$  can be applied to derive an upper bound of  $(\dagger_P)$ . By following the same steps as (4.30) and (4.31), we obtain

$$\int_{\rho}^{\frac{n\rho}{\alpha_1 + \rho}} y^{k-2} dy \leq \begin{cases} \left( \frac{n\rho}{\alpha_1 + \rho} \right)^{k-1}, & \text{if } k \in \mathbb{Z}_{\geq 2}, \\ \log \left( \frac{n}{\alpha_1 + \rho} \right), & \text{if } k = 1, \\ \rho^{k-1}/(1-k), & \text{if } k \in \mathbb{Z}_{k \leq 0}, \end{cases}$$

as a counterpart of the integral in (4.31). By following the same steps as (4.32) and (4.33), we have

$$(\dagger_P) \leq \begin{cases} \frac{\Gamma(n-k+1)}{\Gamma(n-1)} \frac{e n^{k-1}}{k-1}, & \text{if } k \in \mathbb{Z}_{\geq 2}, \\ (n-1) \log \left( \frac{n}{\alpha_1 + \rho} \right), & \text{if } k = 1, \\ \frac{\Gamma(n-k+1)}{\Gamma(n-1)} \frac{e}{(1-k)(\alpha_1 + \rho)^{1-k}}, & \text{if } k \in \mathbb{Z}_{k \leq 0}. \end{cases} \quad (4.36)$$

Note that the probability term in  $(\diamond_P)$  is the same as the CDF of the Erlang( $n-1, n\alpha_1$ ) with rate  $n\alpha_1$  evaluated at  $\frac{1}{n}$  since  $\hat{\alpha}_{1,n} \sim \text{InvG}(n-1, n\alpha_1)$  from (4.5). Thus, we have

$$\begin{aligned} (\diamond_P) &= \frac{1}{1 - C_{1,n}} \frac{1}{G_k(1/n; n)} \frac{\gamma(n-1, \alpha_1)}{\Gamma(n-1)} \\ &\leq \frac{e^{\alpha_1 + \rho}}{1 - C_{1,n}} \frac{\Gamma(n-k+1)}{(\alpha_1 + \rho)^{n-k}} \frac{\gamma(n-1, \alpha_1)}{\Gamma(n-1)} \quad \text{by (4.29)} \\ &\leq \frac{e^{\alpha_1 + \rho}}{1 - C_{1,n}} \frac{\Gamma(n-k+1)}{\Gamma(n-1)} \frac{\alpha_1^{n-1}}{(\alpha_1 + \rho)^{n-k}} \end{aligned} \quad (4.37)$$

$$\begin{aligned} &\leq \frac{e^{\alpha_1 + \rho}}{1 - C_{1,n}} \frac{\Gamma(n-k+1)}{\Gamma(n-1)} \frac{1}{(\alpha_1 + \rho)^{1-k}} \\ &= \mathcal{O}(n^{2-k}), \end{aligned} \quad (4.38)$$

where (4.37) holds from  $\gamma(s, x) \leq x^s$  for any  $s \geq 1$  and  $x > 0$ . Therefore, by combining (4.36) and (4.38) with (4.35) and  $\mathbb{P}[\hat{\sigma} \in K(\epsilon)] = \mathcal{O}(e^{-n\epsilon})$ , we have

$$(\star'_P) \leq \begin{cases} \mathcal{O}(ne^{-n\epsilon}), & \text{if } k \in \mathbb{Z}_{\geq 2} \\ \mathcal{O}(n \log(n)e^{-n\epsilon}), & \text{if } k = 1, \\ \mathcal{O}(n^{2-k}e^{-n\epsilon}), & \text{if } k \in \mathbb{Z}_{\leq 0}. \end{cases} \quad (4.39)$$

By combining (4.39) with (4.34) and (4.27), we obtain for  $k \in \mathbb{Z}_{\geq 0}$

$$\begin{aligned} &\mathbb{E} \left[ \sum_{t=Kn_0+1}^T \mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}^c(\epsilon), \mathcal{M}_\epsilon^c(t)] \right] \\ &\leq \sum_{n=n_0}^T \left( \frac{e^{-n}}{1 - e^{-n}} + \frac{e^{-n\frac{3\epsilon}{4\mu_1}} + \frac{1}{2}e^{-nc\mu_1(\epsilon)}}{C_2(\mu, \epsilon, k)} + (\star'_P) \right) \\ &\leq \sum_{n=n_0}^T \left( \mathcal{O}(e^{-n}) + \mathcal{O}(e^{-n\epsilon}) + \mathcal{O}(e^{-n\epsilon^2}) + \mathcal{O}(\psi(n, k)e^{-n\epsilon}) \right) \\ &= \mathcal{O}(1) + \mathcal{O}(\epsilon^{-1}) + \mathcal{O}(\epsilon^{-2}) + \mathcal{O}(\epsilon^{-\max(2, 3-k)}), \end{aligned}$$

where

$$\psi(n, k) = n\mathbb{1}[k \geq 2] + n \log(n)\mathbb{1}[k = 1] + n^{2-k}\mathbb{1}[k \leq 0].$$

Note that the analysis on term  $(\star'_P)$  also holds for TS-T with  $k \in \mathbb{Z}_{<0}$ . However, differently from the case of  $k \in \{0, 1\}$ , priors with  $k \in \mathbb{Z}_{<0}$  have additional problems in Lemma 4.16 under the event  $\{\hat{\sigma}_{1,n} \in K(\epsilon), \bar{\alpha}_1(n) \leq \alpha_1 + \rho\}$ , where the upper bound becomes a constant  $\frac{1}{2}$ .  $\square$

As a preliminary to the proof of Lemma 4.10, we provide an inequality on the posterior probability that the sampled mean is smaller than a given value, whose proof is given in Section 4.6.8

**Lemma 4.18.** *For any  $i \in [K]$  and  $t \in \mathbb{N}$ , where  $N_i(t) = n$ , it holds for any positive  $\xi \leq \frac{y}{y - \sigma_i}$  and  $k \in \mathbb{Z}$  that*

$$\begin{aligned} &\mathbb{1}[\hat{\sigma}_{i,n}(n) \leq y] \mathbb{P}[\tilde{\mu}_i(t) \leq y | \theta_{i,n}] \\ &\leq \int_{\xi}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx + \left( \frac{y}{\mu((\sigma_i, \xi))} \right)^n \int_1^{\xi} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx, \end{aligned}$$

where  $f_{s,\beta}^{\text{Er}}(\cdot)$  denotes the probability density function of the Erlang distribution with shape  $s \in \mathbb{N}$  and rate  $\beta > 0$ .

The following fact shows another relation between the Erlang distribution and the exponential distribution.

**Fact 4.19.** Let  $x_1, \dots, x_n$  be identically independently distributed with the exponential distribution with the rate parameter  $\alpha$ , i.e.,  $x_s \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\alpha)$  for any  $s \in [n]$ . Then, their sum follows the Erlang distribution with the shape parameter  $n \in \mathbb{N}$  and rate parameter  $\alpha$ , i.e.,  $\sum_{s=1}^n x_s \sim \text{Erlang}(n, \alpha)$ .

*Proof of Lemma 4.10.* When the scale parameter  $\sigma$  is known, the Pareto distribution belongs to the one-dimensional exponential family, where the optimality of TS with the Jeffreys prior was proven by Korda et al. [2013]. Note that the reference prior coincides with the Jeffreys prior for the one-dimensional distributions. In such cases, the posterior on the shape parameter  $\alpha^{\text{one}} > 0$  after observing  $n = N_1(t)$  rewards is given for  $k \in \mathbb{Z}$

$$\alpha^{\text{one}} \mid T(X_n) \sim \text{Erlang}(n - k + 1, q'_n), \quad (4.40)$$

where  $q'_n = \sum_{s=1}^n \log(x_{1,s}) - n \log(\sigma_1)$ . Note that  $q'_n \sim \text{Erlang}(n, \alpha_1)$  from Facts 4.15 and 4.19. Let  $\tilde{\alpha}_1^{\text{one}}$  be a sample from the posterior distribution in (4.40). Then, for one-dimensional Pareto bandits, it holds from (4.9) that

$$\mathbb{P}[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon \mid T(X_{1,n})] = \mathbb{P}_t[\tilde{\alpha}_1^{\text{one}} \geq \beta] = \frac{\Gamma(n - k + 1, \beta q'_n)}{\Gamma(n - k + 1)},$$

where we denoted  $\beta = \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \sigma_1}$  satisfying  $\mu(\sigma_1, \beta) = \mu_1 - \epsilon$ . Therefore, the known optimal result by Korda et al. [2013] (given in Lemma 4.22) can be written as

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}[\mathbb{1}[i(t) \neq 1, \mathcal{M}_\epsilon^c(t)]] &= \sum_{t=1}^T \sum_{n=1}^T \mathbb{E}[\mathbb{1}[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n]] \\ &= \sum_{t=1}^T \sum_{n=1}^T \mathbb{E}[\mathbb{P}_t[i(t) \neq 1, \mathcal{M}_\epsilon^c(t), N_1(t) = n]] \\ &= \sum_{t=1}^T \sum_{n=1}^T \int_0^\infty \frac{\Gamma(n+1, \beta x)}{\Gamma(n+1)} \frac{\alpha_1^n}{\Gamma(n)} x^{n-1} e^{-\alpha_1 x} dx \quad (4.41) \\ &\leq \mathcal{O}(\epsilon^{-1}), \end{aligned}$$

where we injected the density function of the Erlang distribution into the last equality.

On the other hand, for two-parameter Pareto bandits where the scale parameter is unknown, it holds by the law of total expectation that

$$\begin{aligned} \mathbb{E}[\mathbb{1}[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t)]] &= \mathbb{E}_{\hat{\theta}_{1,n}}[\mathbb{P}_t[i(t) \neq 1, \mathcal{B}_{1,N_1(t)}(\epsilon), \mathcal{M}_\epsilon^c(t)]] \\ &= \mathbb{E}_{\hat{\theta}_{1,n}}[\mathbb{1}[\mathcal{B}_{1,N_1(t)}(\epsilon)] \mathbb{P}_t[i(t) \neq 1, \mathcal{M}_\epsilon^c(t)]], \end{aligned}$$

where the last equality holds since  $\mathcal{B}$  is determined by  $T(X_{1,n})$ .

From Lemma 4.18 with  $y = \mu_1 - \epsilon$ , it holds for any  $\xi \leq \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \sigma_1} = \beta$  that

$$\begin{aligned} \mathbb{1}[\mathcal{B}_{1,n}(\epsilon)] \mathbb{P}_t[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon] &\leq \mathbb{1}[\mathcal{B}_{1,n}(\epsilon)] \left( \left( \frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))} \right)^n \int_1^\xi f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx + \int_\xi^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx \right) \\ &\leq \mathbb{1}[\mathcal{B}_{1,n}(\epsilon)] \left( \left( \frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))} \right)^n \left( 1 - \frac{\Gamma(n-k, \frac{n}{\alpha_{1,n}} \xi)}{\Gamma(n-k)} \right) + \frac{\Gamma(n-k, \frac{n}{\alpha_{1,n}} \xi)}{\Gamma(n-k)} \right) \quad (4.42) \end{aligned}$$



which is a convex combination of 1 and  $\left(\frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))}\right)^n$ . Therefore, RHS of (4.42) increases

as  $\frac{\Gamma\left(n-k, \frac{n}{\alpha_{1,n}}\xi\right)}{\Gamma(n-k)}$  increases. From the definition of  $\Gamma(n, x)$ , it holds that  $\Gamma(n, x) \geq \Gamma(n, x + y)$  for any positive  $y > 0$  and  $\Gamma(n+1, x) = n\Gamma(n, x) + x^n e^{-x}$ . Since  $\frac{n}{\hat{\alpha}_{1,n}} \leq \frac{n}{\alpha_{1,n}}$  holds for any  $n \in \mathbb{N}$ , it holds for  $k \in \mathbb{Z}_{\geq 0}$  that

$$\frac{\Gamma\left(n-k, \frac{n}{\hat{\alpha}_{1,n}}\xi\right)}{\Gamma(n-k)} \leq \frac{\Gamma\left(n-k, \frac{n}{\alpha_{1,n}}\xi\right)}{\Gamma(n-k)} \leq \frac{\Gamma\left(n, \frac{n}{\alpha_{1,n}}\xi\right)}{\Gamma(n)}.$$

Let us denote  $Y_n := \frac{n}{\hat{\alpha}_{1,n}} = \sum_{i=1}^n \log(r_{1,s}) - n \log(\hat{\sigma}_{1,n})$ , which follows the Erlang distribution with shape  $n-1$  and rate  $\alpha_1$  [Malik, 1970]. By taking expectation with respect to  $\hat{\sigma}_{1,n}$ , we have for any  $\xi \leq \beta$  that

$$\begin{aligned} & \mathbb{E}_{\hat{\sigma}_{1,n}}[\mathbb{1}[\mathcal{B}_{1,n}(\epsilon)]\mathbb{P}_t[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon]] \\ & \leq \int_{\sigma_1}^{\sigma_1 + \epsilon} \left( \left( \frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))} \right)^n \left( 1 - \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)} \right) + \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)} \right) \mathbb{P}[\hat{\sigma}_{1,n} = x] dx \\ & = \mathbb{P}[\mathcal{B}_{1,n}(\epsilon)] \left( \left( \frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))} \right)^n \left( 1 - \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)} \right) + \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)} \right) \\ & = \left( 1 - \left( \frac{\sigma_1}{\sigma_1 + \epsilon} \right)^{n\alpha_1} \right) \left( \left( \frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))} \right)^n \left( 1 - \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)} \right) + \frac{\Gamma(n, \xi Y_n)}{\Gamma(n)} \right), \end{aligned}$$

where we used  $\hat{\sigma}_{1,n} \sim \text{Pareto}(\sigma_1, n\alpha_1)$  in (4.5) for the last equation.

Therefore, under the two-parameter Pareto distribution, the following holds for any  $\xi \leq \beta$  under both TS and TS-T with  $k \geq 0$  that

$$\begin{aligned} & \mathbb{E}_{\hat{\sigma}_{1,n}, \hat{\alpha}_1}[\mathbb{1}[\mathcal{B}_{1,n}(\epsilon)]\mathbb{P}[\tilde{\mu}_1(t) \leq \mu_1 - \epsilon | T(X_{1,n})]] \\ & \leq \left( 1 - \left( \frac{\sigma_1}{\sigma_1 + \epsilon} \right)^{n\alpha_1} \right) \int_0^\infty \left( \left( \frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))} \right)^n \left( 1 - \frac{\Gamma(n, \xi x)}{\Gamma(n)} \right) + \frac{\Gamma(n, \xi x)}{\Gamma(n)} \right) \\ & \quad \cdot \frac{\alpha_1^{n-1}}{\Gamma(n-1)} x^{n-2} e^{-\alpha_1 x} dx. \end{aligned} \quad (4.43)$$

Notice that for sufficiently large  $n$ , the main integrand of the RHS of (4.43) is

$$\frac{\Gamma(n, \xi x)}{\Gamma(n)} \frac{\alpha_1^{n-1}}{\Gamma(n-1)} x^{n-2} e^{-\alpha_1 x}.$$

Recall the integrand appearing in TS for the one-parameter Pareto models in (4.41). From  $\Gamma(n+1, x) = n\Gamma(n, x) + x^n e^{-x}$ , we have

$$\begin{aligned} \frac{\Gamma(n+1, \beta x)}{\Gamma(n+1)} \frac{\alpha_1^n}{\Gamma(n)} x^{n-1} e^{-\alpha_1 x} &= \frac{n\Gamma(n, \beta x) + (\beta x)^n e^{-\beta x}}{\Gamma(n+1)} \frac{\alpha_1^n}{\Gamma(n)} x^{n-1} e^{-\alpha_1 x} \\ &= \frac{\Gamma(n, \beta x)}{\Gamma(n)} \frac{\alpha_1^n}{\Gamma(n)} x^{n-1} e^{-\alpha_1 x} + \frac{(\alpha_1 \beta)^n x^{2n-1} e^{-(\alpha_1 + \beta)x}}{\Gamma(n+1)\Gamma(n)}. \end{aligned}$$

Therefore, for every  $n$ , one can carefully choose  $\xi \leq \beta$  such that

$$\begin{aligned} & \frac{\Gamma(n, \xi x)}{\Gamma(n)} \frac{\alpha_1^{n-1}}{\Gamma(n-1)} x^{n-2} e^{-\alpha_1 x} \\ & \leq \frac{\Gamma(n, \beta x)}{\Gamma(n)} \frac{\alpha_1^n}{\Gamma(n)} x^{n-1} e^{-\alpha_1 x} + \frac{(\alpha_1 \beta)^n x^{2n-1} e^{-(\alpha_1 + \beta)x}}{\Gamma(n+1)\Gamma(n)}, \end{aligned}$$

which concludes the proof.

Note that from  $\Gamma(n, x) \geq \Gamma(n, x + y)$  for any positive  $x, y > 0$  and  $\xi' \leq \beta$ , we have for any  $x > 0$  that

$$\frac{\Gamma(n+1, \xi'x)}{\Gamma(n+1)} \geq \frac{\Gamma(n+1, \beta x)}{\Gamma(n+1)}.$$

Therefore, for  $k \in \mathbb{Z}_{\leq -1}$ , we might not be able to apply the results by Korda et al. [2013].  $\square$

#### 4.6.8 Proofs of technical lemmas on fundamental inequalities

This section presents detailed proofs of Lemmas 4.16 and 4.18, which evaluates the posterior distribution of the mean for TS and TS-T.

*Proof of Lemma 4.16.* Similarly to other proofs, fix  $t$  and let  $N_1(t) = n$ . To simplify notations, we drop the argument  $t$  of  $\tilde{\sigma}_1(t)$ ,  $\tilde{\alpha}_1(t)$  and  $\tilde{\mu}_1(t)$ . Recall that this lemma is based on  $\alpha_{1,n}$  notation.

**(1) On  $\{\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon\}$**

Under the condition  $\{\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon\}$ , the event  $\{\tilde{\mu}_1 \leq \mu_1 - \epsilon\}$  is eventually determined by the value of  $\tilde{\sigma}_1$  since  $\{\tilde{\sigma}_1 \in (\mu_1 - \epsilon, \hat{\sigma}_{1,n}]\}$  is a sufficient condition to  $\{\tilde{\mu}_1 > \mu_1 - \epsilon\}$ . Therefore, if  $\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon$ , then

$$\begin{aligned} p_n(\epsilon|\theta_{1,n}) &= \int_1^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) \mathbb{P}\left[\tilde{\sigma}_1 \leq (\mu_1 - \epsilon) \frac{x-1}{x} \mid \tilde{\alpha} = x\right] dx \\ &= \int_1^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) \left(\frac{\mu_1 - \epsilon}{\hat{\sigma}_{1,n}} \frac{x-1}{x}\right)^{nx} dx. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{1}[\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon] p_n(\epsilon|\theta_{1,n}) &= \mathbb{1}[\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon] \int_1^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) \left(\frac{\mu_1 - \epsilon}{\hat{\sigma}_{1,n}} \frac{x-1}{x}\right)^{nx} dx \\ &\leq \mathbb{1}[\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon] \int_1^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) \left(1 - \frac{1}{x}\right)^{nx} dx \\ &\leq \mathbb{1}[\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon] \int_1^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) e^{-n} dx \\ &\leq \mathbb{1}[\hat{\sigma}_{1,n} \geq \mu_1 - \epsilon] e^{-n}, \end{aligned}$$

where the second inequality holds from  $\mu_1 - \epsilon \leq \hat{\sigma}_{1,n}$ .

**(2) On  $\{\hat{\sigma}_{1,n} \in K(\epsilon), \alpha_{1,n} \leq \alpha_1 + \rho\}$**

By applying Lemma 4.18 with  $y = \mu_1 - \epsilon$ , we have for any  $\xi \leq \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon - \sigma_1}$  that

$$\begin{aligned} \mathbb{1}[\hat{\sigma}_{1,n} < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha + \rho] p_n(\epsilon|\theta_{1,n}) &\leq \left(\frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))}\right)^n \int_1^\xi f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx + \int_\xi^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx \\ &\leq \left(\frac{\mu_1 - \epsilon}{\mu((\sigma_1, \xi))}\right)^n \int_0^\xi f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx + \int_\xi^\infty f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx. \quad (4.44) \end{aligned}$$

Let us define  $\bar{\rho} := \rho_\theta(\epsilon/2)$ . Then, it satisfies  $\mu(\sigma_1, \alpha_1 + \bar{\rho}) = \mu_1 - \frac{\epsilon}{4}$  and

$$\alpha_1 + \bar{\rho} = \frac{\mu - \epsilon/4}{\mu - \epsilon/4 - \sigma_1} < \frac{\mu - \epsilon}{\mu - \epsilon - \sigma_1},$$

where the inequality holds from the decreasing property of the function  $\frac{x}{x-y}$  with respect to  $x > y$ . By replacing  $\xi$  with  $\alpha_1 + \bar{\rho}$  in (4.44), we have

$$\begin{aligned}
& \mathbb{1}[\hat{\sigma}_{1,n} < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho] p_n(\epsilon | \bar{\theta}_{1,n}) \\
& \leq \mathbb{1}[\hat{\sigma}_{1,n} < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho] \\
& \quad \cdot \left( \left( \frac{\mu_1 - \epsilon}{\mu_1 - \epsilon/4} \right)^n \int_0^\xi f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx + \int_{\alpha_1 + \bar{\rho}}^\infty f_{n-k, \frac{n}{\alpha_{1,n}(\bar{\rho})}}^{\text{Er}}(x) dx \right) \\
& \leq \mathbb{1}[\hat{\sigma}_{1,n} < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho] \\
& \quad \cdot \left( e^{-n \left( \frac{3\epsilon}{4\mu_1 - \epsilon} \right)} (1 - \mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}]) + \mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}] \right). \tag{4.45}
\end{aligned}$$

Let  $Z_n$  be a random variable that follows the chi-squared distribution with  $n$  degree of freedom and  $F_n^{\chi^2}(\cdot)$  denote the CDF of  $Z_n$ . Then, it holds that

$$\begin{aligned}
& \mathbb{P} \left[ \tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho \right] \\
& = \mathbb{P} \left[ Z \geq \frac{2n}{\alpha_{1,n}} (\alpha_1 + \bar{\rho}), \alpha_{1,n} \leq \alpha_1 + \rho \right] \quad \because \text{Fact 4.11} \\
& \leq \mathbb{P} \left[ Z \geq 2n \frac{\alpha_1 + \bar{\rho}}{\alpha_1 + \rho} \right] \\
& \leq \mathbb{P} \left[ Z \geq 2n \frac{\mu_1 - \epsilon/4}{\mu_1 - \epsilon/2} \right] \\
& = 1 - F_{2n-2k}^{\chi^2}(2n(1 + \zeta)),
\end{aligned}$$

where  $\zeta = \frac{\epsilon}{4\mu_1 - 2\epsilon} \in (0, 1)$ . By applying Lemma 4.23, we have if  $n\zeta > -k$ ,

$$\begin{aligned}
& \mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] \\
& \leq 1 - F_{2n-2k}^{\chi^2}(2n(1 + \zeta)) \\
& < \frac{1}{2} \frac{\sqrt{2\pi}(n-k)^{n-k-1/2} e^{-(n-k)}}{\Gamma(n-k)} \\
& \quad \cdot \text{erfc} \left( \sqrt{n(\zeta + k) - (n-k) \log \frac{n(1+\zeta)}{n-k}} \right),
\end{aligned}$$

where  $\text{erfc}(\cdot)$  denotes the complementary error function. For  $n \geq 1/2$ , it holds from Stirling's formula that

$$\sqrt{2\pi} n^{n-1/2} e^{-n} \leq \Gamma(n) \leq \sqrt{2\pi} e^{1/6} n^{n-1/2} e^{-n},$$

which results in

$$\mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] < \frac{1}{2} \text{erfc} \left( \sqrt{n(\zeta + k) - (n-k) \log \frac{n(1+\zeta)}{n-k}} \right). \tag{4.46}$$

Notice that  $(n-k) \log \frac{n(1+\zeta)}{n-k} > 0$  always holds from the assumption of  $n\zeta > -k$  where priors with  $k \in \mathbb{Z}_{\geq 0}$  satisfies regardless of  $n$ . Thus, if  $\zeta + k \leq 0$ , then the argument in the complementary error function in (4.46) becomes negative. This makes the upper bound in (4.46) greater than or equal to  $\frac{1}{2}$ . Therefore, for the priors with  $k \in \mathbb{Z}_{< 0}$ , the right term in (4.46) is bounded by a constant since  $\zeta \in (0, 1)$ .

Since the complementary error function is a decreasing function, for priors with  $k \in \mathbb{Z}_{\geq 0}$ , it holds from (4.46) that

$$\mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] \leq \frac{1}{2} \text{erfc} \left( \sqrt{n(\zeta - \log(1 + \zeta))} \right),$$

where we substitute  $k = 0$ . By the change of variables, the complementary error function is bounded for any  $x \geq 0$  as follows [Simon and Divsalar, 1998]:

$$\operatorname{erfc}(x) \leq e^{-x^2},$$

which implies

$$\mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \bar{\rho}, \alpha_{1,n} \leq \alpha_1 + \rho] \leq \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}, \quad (4.47)$$

where  $c_{\mu_1}(\epsilon) = \zeta - \log(1 + \zeta) > 0$  is a deterministic constants on  $\mu_1$  and  $\epsilon$ . By combining (4.47) with (4.45), we have

$$\begin{aligned} \mathbb{1}[\hat{\sigma}_{1,n} < \mu_1 - \epsilon, \alpha_{1,n} \leq \alpha_1 + \rho] p_n(\epsilon | \theta_{1,n}) \\ \leq e^{-n\frac{3\epsilon}{4\mu_1}} \left(1 - \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}\right) + \frac{1}{2}e^{-nc_{\mu_1}(\epsilon)}. \end{aligned}$$

Denote the RHS by  $h(\mu_1, \epsilon, n)$  concludes the proof.

From the power-series expansion of  $\log(1 + x)$ , we have  $\log(1 + x) \geq x - \frac{x^2}{2} + \frac{x^3}{3}$  for  $x \in (0, 1)$  and

$$\begin{aligned} c_{\mu_1}(\epsilon) = \zeta - \log(1 + \zeta) &\leq \frac{\zeta^2}{2} - \frac{\zeta^3}{3} = \frac{\zeta^2}{6}(3 - 2\zeta) \\ &\leq \frac{\zeta^2}{2} = \mathcal{O}(\epsilon^{-2}). \end{aligned}$$

**(3) On  $\{\hat{\sigma}_{1,n} \in K(\epsilon), \alpha_{1,n} \geq \alpha_1 + \rho\}$**

By applying Lemma 4.18 with  $y = \mu_1 - \epsilon$  and  $\xi = \alpha_1 + \rho$ , we have

$$\begin{aligned} \mathbb{1}[\hat{\sigma}_1 < \mu_1 - \epsilon] p_n(\epsilon | \theta_{1,n}) \\ \leq \left(\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon/2}\right)^n \int_1^{\alpha_1 + \rho} f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx + \int_{\alpha_1 + \rho}^{\infty} f_{n-k, \frac{n}{\alpha_{1,n}}}^{\text{Er}}(x) dx \\ = C_1(\mu_1, \epsilon, n) \mathbb{P}[\tilde{\alpha}_1 \in [1, \alpha_1 + \rho] | \alpha_{1,n}] + \mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \rho | \alpha_{1,n}] \\ \leq C_1(\mu_1, \epsilon, n) \mathbb{P}[\tilde{\alpha}_1 \leq \alpha_1 + \rho | \alpha_{1,n}] + \mathbb{P}[\tilde{\alpha}_1 \geq \alpha_1 + \rho | \alpha_{1,n}], \end{aligned}$$

where  $C_1(\mu_1, \epsilon, n) := \left(\frac{\mu_1 - \epsilon}{\mu_1 - \epsilon/2}\right)^n \leq e^{-n\frac{\epsilon}{2\mu_1 - \epsilon}} < 1$ . Since  $\tilde{\alpha}_1$  follows the Erlang distribution with shape  $n - k$  and rate  $\frac{n}{\alpha_{1,n}}$ , it holds that

$$\mathbb{P}[\tilde{\alpha}_1 \leq \alpha_1 + \rho | \alpha_{1,n}] = \frac{\gamma\left(n - k, \frac{n(\alpha_1 + \rho)}{\alpha_{1,n}}\right)}{\Gamma(n - k)},$$

where  $\gamma(\cdot, \cdot)$  denotes the lower incomplete gamma function. Therefore, letting

$$G_k(x; n) := \frac{\gamma(n - k, n(\alpha_1 + \rho)x)}{\Gamma(n - k)}$$

concludes the proof.  $\square$

*Proof of Lemma 4.18.* Fix a time index  $t$  with  $N_i(t) = n$  and denote  $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot | T(X_{i,n})]$ . To simplify notations, we drop the argument  $t$  of  $\tilde{\sigma}_i(t)$ ,  $\tilde{\alpha}_i(t)$ , and  $\tilde{\mu}_i(t)$ .

When  $\hat{\sigma}_{i,n} < y$  holds,  $\tilde{\mu}_i \leq y$  can hold regardless of the value of  $\tilde{\sigma}_i$  if  $\hat{\sigma}_{i,n} \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i - 1} \leq y$  holds since  $\tilde{\sigma}_i \in (0, \hat{\sigma}_{i,n}]$  holds from its posterior distribution in (4.7). Hence, if  $\hat{\sigma}_{i,n} < y$ , then

$$\tilde{\alpha}_i \geq \frac{y}{y - \hat{\sigma}_{i,n}} \implies \tilde{\mu}_i \leq y. \quad (4.48)$$

When  $1 < \tilde{\alpha}_i < \frac{y}{y - \tilde{\sigma}_{i,n}}$ ,

$$\tilde{\mu}_i = \tilde{\sigma}_i \frac{\tilde{\alpha}_i}{\tilde{\alpha}_i - 1} \leq y \Leftrightarrow \tilde{\sigma}_i \leq y \frac{\tilde{\alpha}_i - 1}{\tilde{\alpha}_i}. \quad (4.49)$$

Since  $\tilde{\alpha}_i \leq 1$  implies  $\tilde{\mu}_i = \infty$ , from (4.48) and (4.49), it holds that

$$\begin{aligned} \mathbb{P}_t[\tilde{\mu}_i \leq y] &= \int_1^{\frac{y}{y - \tilde{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \mathbb{P}_t \left[ \tilde{\sigma}_i \leq \frac{x-1}{x} y \right] dx + \int_{\frac{y}{y - \tilde{\sigma}_{i,n}}}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \\ &= \int_1^{\frac{y}{y - \tilde{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left( \frac{x-1}{\tilde{\sigma}_{i,n} x} y \right)^{nx} dx + \int_{\frac{y}{y - \tilde{\sigma}_{i,n}}}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx, \end{aligned} \quad (4.50)$$

where we injected the CDF of the conditional posterior of (4.7) in (4.50). Take any finite  $y' > y$  and let  $\xi := \frac{y'}{y' - \sigma_i} < \frac{y}{y - \sigma_i}$  such that  $\mu((\sigma_i, \xi)) = y'$ . Since  $\frac{a}{a-b}$  is decreasing with respect to  $a > b > 0$ , one can see that

$$\begin{aligned} \mathbb{P}_t[\tilde{\mu}_i \leq y] &= \int_1^{\frac{y}{y - \tilde{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left( \frac{x-1}{\hat{\sigma}_{i,n} x} y \right)^{nx} dx + \int_{\frac{y}{y - \tilde{\sigma}_{i,n}}}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \\ &\leq \int_1^{\frac{y'}{y' - \tilde{\sigma}_{i,n}}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left( \frac{x-1}{\hat{\sigma}_{i,n} x} y \right)^{nx} dx + \int_{\frac{y'}{y' - \tilde{\sigma}_{i,n}}}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \\ &\leq \int_1^{\frac{y'}{y' - \sigma_i}} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left( \frac{x-1}{\hat{\sigma}_{i,n} x} y \right)^{nx} dx + \int_{\frac{y'}{y' - \sigma_i}}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \\ &\leq \int_1^{\xi} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) \left( \frac{x-1}{\sigma_i x} y \right)^{nx} dx + \int_{\xi}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \end{aligned} \quad (4.51)$$

$$\begin{aligned} &\leq \left( \frac{\xi-1}{\sigma_i \xi} y \right)^n \int_1^{\xi} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx + \int_{\xi}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx \\ &= \left( \frac{y}{\mu((\sigma, \xi))} \right)^n \int_1^{\xi} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx + \int_{\xi}^{\infty} f_{n-k, \frac{n}{\alpha_{i,n}}}^{\text{Er}}(x) dx, \end{aligned} \quad (4.52)$$

where (4.51) comes from  $\hat{\sigma}_{i,n} \geq \sigma_i$  and we used the increasing property of  $\frac{x-1}{x}$  in (4.52).  $\square$

#### 4.6.9 Proof of suboptimality of TS

As shown in proofs of Lemma 4.9, the integral term in (4.31) diverges for  $k \in \mathbb{Z}_{\leq 1}$  without the restriction on  $\hat{\alpha}$ . In this section, we provide the partial proof of Theorem 4.3 for  $k \in \mathbb{Z}_{\leq 0}$ , which shows the necessity of such requirement to achieve asymptotic optimality.

*Proof of Theorem 4.3.* We consider a two-armed bandit problem with  $\theta_1 = (\sigma_1, \alpha_1)$  and  $\theta_2 = (\sigma_2, \alpha_2)$ . Assume  $1 < \alpha_1 < \alpha_2$  and  $\tilde{\mu}_2(s) = \mu_2 = \sigma_2 \frac{\alpha_2}{\alpha_2 - 1}$  for all  $s \in \mathbb{N}$ . Recall that TS starts from playing every arm twice for priors  $k \leq 1$ , i.e.,  $N_i(s) \geq 2$  holds for all  $i \in \{1, 2\}$ . We have for  $T \geq 5$

$$\begin{aligned} \mathbb{E}[\text{Reg}(T)] &= \Delta_2 \mathbb{E} \left[ \sum_{t=1}^T \mathbb{1}[i(t) = 2] \right] \\ &\geq \Delta_2 \mathbb{E} \left[ \sum_{t=5}^T \mathbb{1}[i(t) = 2, N_1(t) = 2] \right]. \end{aligned}$$

From the definition of  $N_1(\cdot)$ ,  $\{i(s) \neq 2, N_1(s) = 2\}$  implies  $N_1(t) > 2$  for  $t > s$ . Therefore, for any  $t \geq 5$ ,

$$\begin{aligned} \{i(t) = 2, N_1(t) = 2\} &\Leftrightarrow \{\forall s \in [1, t-4] : i(s+4) = 2\} \\ &\Leftrightarrow \{\forall s \in [1, t-4] : \tilde{\mu}_1(s+4) < \mu_2\}. \end{aligned}$$

Let  $T' = T - 4$ , then we have

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=5}^T \mathbb{1}[i(t) = 2, N_1(t) = 2] \right] &= \mathbb{E} \left[ \sum_{t=5}^T \mathbb{1}[\forall s \in [1, t-4] : \tilde{\mu}_1(s+4) < \mu_2] \right] \\ &= \mathbb{E} \left[ \sum_{s=1}^{T'} (\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\sigma}_{1,2}, \hat{\alpha}_{1,2}])^s \right]. \end{aligned} \quad (4.53)$$

Notice that  $\hat{\sigma}_{1, N_1(s)} = \hat{\sigma}_{1,2}$  and  $\hat{\alpha}_{1, N_1(s)} = \hat{\alpha}_{1,2}$  hold for all  $s \geq 2$  since only  $i(s) = 2$  holds for all  $s \geq 2$ .

Here, we first provide the lower bound on  $\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\sigma}_{1,2}, \hat{\alpha}_{1,2}]$ . From (4.50), it holds for  $y \geq \hat{\sigma}_1(n)$  that

$$\begin{aligned} \mathbb{P}_t[\tilde{\mu}_a \leq y] &= \int_1^{\frac{y}{y-\hat{\sigma}_{1,n}}} f_{n-k, \frac{n}{\hat{\alpha}_{1,n}}}^{\text{Er}}(x) \left( \frac{x-1}{\hat{\sigma}_{1,n}x} y \right)^{nx} dx + \int_{\frac{y}{y-\hat{\sigma}_{1,n}}}^{\infty} f_{n-k, \frac{n}{\hat{\alpha}_{1,n}}}^{\text{Er}}(x) dx \quad \text{by (4.50)} \\ &\geq \int_{\frac{y}{y-\hat{\sigma}_{1,n}}}^{\infty} f_{n-k, \frac{n}{\hat{\alpha}_{1,n}}}^{\text{Er}}(x) dx. \end{aligned}$$

By letting  $n = 2$  and  $y = \mu_2$ , we have for  $k \in \mathbb{Z}_{\leq 1}$

$$\begin{aligned} \mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\sigma}_{1,2}, \hat{\alpha}_{1,2}] &\geq \mathbb{1}[\hat{\sigma}_{1,2} \leq \sigma_2] \int_{\frac{\mu_2}{\mu_2-\hat{\sigma}_{1,2}}}^{\infty} f_{2-k, \frac{2}{\hat{\alpha}_{1,2}}}^{\text{Er}}(x) dx \\ &\geq \mathbb{1}[\hat{\sigma}_{1,2} \leq \sigma_2] \int_{\alpha_2}^{\infty} f_{2-k, \frac{2}{\hat{\alpha}_{1,2}}}^{\text{Er}}(x) dx \quad \because \alpha_2 = \frac{\mu_2}{\mu_2 - \sigma_2} \\ &= \mathbb{1}[\hat{\sigma}_{1,2} \leq \sigma_2] \frac{\Gamma(2-k, \frac{2\alpha_2}{\hat{\alpha}_{1,2}})}{\Gamma(2-k)}, \end{aligned} \quad (4.54)$$

where  $\Gamma(\cdot, \cdot)$  is the upper incomplete Gamma function.

**(1) Priors  $k \in \mathbb{Z}_{\leq 1}$**

Note that  $\Gamma(n, x)$  is an increasing function with respect to  $n$  for fixed  $x$ . Therefore, (4.54) implies that if the lower bound of regret for the reference prior is larger than the lower bound, then every prior with  $k \in \mathbb{Z}_{\leq 1}$  is suboptimal. Therefore, let us consider the case  $k = 1$ , where we can rewrite (4.54) as

$$\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\theta}_{1,2}] \geq \mathbb{1}[\hat{\sigma}_{1,2} \leq \sigma_2] \frac{\Gamma(1, \frac{2\alpha_2}{\hat{\alpha}_{1,2}})}{\Gamma(1)} = e^{-\frac{2\alpha_2}{\hat{\alpha}_{1,2}}}. \quad (4.55)$$

Since  $\hat{\alpha}_{1,2} \sim \text{InvG}(1, 2\alpha_1)$  in (4.5),  $z := \frac{2\alpha_2}{\hat{\alpha}_{1,2}}$  follows an exponential distribution with rate parameter  $\alpha_1/\alpha_2 < 1$ , i.e.,  $z \sim \text{Exp}(\alpha_1/\alpha_2)$ . By combining (4.55) with (4.53), we have

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=5}^T \mathbb{1}[i(t) = 2, N_1(t) = 2] \right] &\geq \mathbb{E}_{\hat{\sigma}, z} \left[ \sum_{s=1}^{T'} \left( \mathbb{1}[\hat{\sigma}_{1,2} \leq \sigma_2] e^{-z} \right)^s \right] \\ &= \mathbb{P}[\hat{\sigma}_{1,2} \leq \sigma_2] \mathbb{E}_{z \sim \text{Exp}(\alpha_1/\alpha_2)} \left[ \sum_{s=1}^{T'} e^{-zs} \right], \end{aligned} \quad (4.56)$$

where we used the stochastic independence of  $\hat{\alpha}$  and  $\hat{\sigma}$ . Here,

$$\begin{aligned}
\mathbb{E}_{z \sim \text{Exp}(\alpha_1/\alpha_2)} \left[ \sum_{s=1}^{T'} e^{-zs} \right] &= \mathbb{E}_{z \sim \text{Exp}(\alpha_1/\alpha_2)} \left[ (1 - e^{-zT'}) \frac{e^{-z}}{1 - e^{-z}} \right] \\
&= \int_0^\infty (1 - e^{-xT'}) \frac{e^{-x}}{1 - e^{-x}} e^{-\frac{\alpha_1}{\alpha_2} x} dx \\
&\geq \int_0^\infty (1 - e^{-xT'}) \frac{e^{-2x}}{1 - e^{-x}} dx \quad \text{by } \frac{\alpha_1}{\alpha_2} < 1 \\
&\geq \left(1 - \frac{1}{e}\right) \int_{\frac{1}{T'}}^\infty \frac{e^{-2x}}{1 - e^{-x}} dx \\
&= \left(1 - \frac{1}{e}\right) [\log(e^x - 1) - x + e^{-x}]_{x=\frac{1}{T'}}^\infty \\
&\geq \left(1 - \frac{1}{e}\right) \left( \log T' + 1 - \frac{3}{2T'} \right), \tag{4.57}
\end{aligned}$$

where the last inequality holds from its power series expansion

$$\log(e^x - 1) - x + e^{-x} \geq \log(x) + 1 - \frac{3}{2}x$$

and  $\lim_{x \rightarrow \infty} \log(e^x - 1) - x + e^{-x} = 0$ . By combining (4.57) with (4.56) and (4.53) and elementary calculation with  $\hat{\sigma}_{1,2} \sim \text{Pareto}(\sigma_1, 2\alpha_1)$ , we have

$$\begin{aligned}
\mathbb{E}[\text{Reg}(T)] &\geq \Delta_2 \left(1 - \left(\frac{\sigma_1}{\sigma_2}\right)^{2\alpha_1}\right) \left(1 - \frac{1}{e}\right) \left(\log T' + 1 - \frac{3}{2T'}\right) \\
&= \Delta_2 \left(1 - \left(\frac{\sigma_1}{\sigma_2}\right)^{2\alpha_1}\right) \left(1 - \frac{1}{e}\right) \left(\log(T + 4) + 1 - \frac{3}{2(T + 4)}\right).
\end{aligned}$$

Therefore, under TS with  $k \in \mathbb{Z}_{\leq 1}$ , there exists a constant  $\xi^{\text{Pa}}$  independent of  $\alpha_2$  such that

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\log T} \geq \Delta_2 \xi^{\text{Pareto}}.$$

## (2) Priors $k \in \mathbb{Z}_{\leq 0}$

Similarly to the case of  $k \in \mathbb{Z}_{\leq 1}$ , it is sufficient to consider the case  $k = 0$ , where we can rewrite (4.54) as

$$\mathbb{P}[\tilde{\mu}_1 \leq \mu_2 | \hat{\theta}_{1,2}] \geq \mathbb{1}[\hat{\sigma}_{1,2} \leq \sigma_2] \frac{\Gamma\left(2, \frac{2\alpha_2}{\hat{\alpha}_{1,2}}\right)}{\Gamma(2)}.$$

From the definition of the upper incomplete Gamma function, we have

$$g(z) := \Gamma(2, z) = \int_z^\infty x^1 e^{-x} dx = e^{-z}(z + 1),$$

as a counterpart of  $e^{-z}$  in (4.56) with the same notations  $z = \frac{2\alpha_2}{\hat{\alpha}_{1,2}} \sim \text{Exp}\left(\frac{\alpha_1}{\alpha_2}\right)$ . Since  $g(z) \geq 1 - z^2$  holds for  $z \in [0, 1]$ , one can obtain by replacing  $e^{-zs}$  in (4.56) with  $g(z)^s$  that

$$\begin{aligned}
\mathbb{E}_z \left[ \sum_{s=1}^{T'} (g(z))^s \right] &\geq \mathbb{E}_z \left[ \mathbb{1}[z \in (0, 1)] \sum_{s=1}^{T'} (1 - z^2)^s \right] \\
&= \mathbb{E}_z \left[ \mathbb{1}[z \in (0, 1)] (1 - (1 - z^2)^{T'}) \frac{1 - z^2}{z^2} \right].
\end{aligned}$$

Since  $z \in \left(0, \frac{1}{\sqrt{T'}}\right]$ ,  $(1 - z^2)^{T'} \leq \frac{1}{1 + T'z^2}$  holds, we have  $1 - (1 - z^2)^{T'} \geq \frac{T'z^2}{1 + T'z^2}$ . By applying this fact, we have for  $T' > 1$ ,

$$\begin{aligned} \mathbb{E}_z \left[ \sum_{s=1}^{T'} (g(z))^s \right] &\geq \mathbb{E}_z \left[ \frac{T'(1 - z^2)}{1 + T'z^2} \mathbb{1} \left[ z \in \left(0, \frac{1}{\sqrt{T'}}\right] \right] \right] \\ &\geq \mathbb{E}_{z \sim \text{Exp}(\alpha_1/\alpha_2)} \left[ \left( \frac{T'}{2} - \frac{1}{2} \right) \mathbb{1} \left[ z \in \left(0, \frac{1}{\sqrt{T'}}\right] \right] \right] \\ &= \left( \frac{T'}{2} - \frac{1}{2} \right) \int_0^{\frac{1}{\sqrt{T'}}} \frac{\alpha_1}{\alpha_2} e^{-\frac{\alpha_1}{\alpha_2} z} dz \\ &= \left( \frac{T'}{2} - \frac{1}{2} \right) \left( 1 - e^{-\frac{\alpha_1}{\alpha_2 \sqrt{T'}}} \right). \end{aligned}$$

Notice that  $e^{-x} \leq 1 - \frac{x}{2}$  holds for  $x < 1$ , which gives

$$\begin{aligned} \mathbb{E}_z \left[ \sum_{s=1}^{T'} (g(z))^s \right] &\geq \left( \frac{T'}{2} - \frac{1}{2} \right) \left( 1 - e^{-\frac{\alpha_1}{\alpha_2 \sqrt{T'}}} \right) \\ &\geq \left( \frac{T'}{2} - \frac{1}{2} \right) \frac{\alpha_1}{2\alpha_2 \sqrt{T'}} = \frac{\alpha_1}{4\alpha_2} \left( \sqrt{T'} - \frac{1}{\sqrt{T'}} \right). \end{aligned} \quad (4.58)$$

By applying (4.58) to (4.53), we obtain for  $k \in \mathbb{Z}_{\leq 0}$  and  $T' = T - 4 > 1$ ,

$$\begin{aligned} \mathbb{E}[\text{Reg}(T)] &\geq \Delta_2 \frac{\alpha_1}{4\alpha_2} \left( 1 - \left( \frac{\sigma_1}{\sigma_2} \right)^{2\alpha_1} \right) \left( \sqrt{T'} - \frac{1}{\sqrt{T'}} \right) \\ &= \mathcal{O}(\sqrt{T}). \end{aligned}$$

Therefore, under TS with priors  $k \in \mathbb{Z}_{\leq 0}$ , there exists a constant  $\xi^{\text{Pa}'} > 0$  independent of  $\alpha_2$  such that

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[\text{Reg}(T)]}{\sqrt{T}} \geq \Delta_2 \xi^{\text{Pa}'}.$$

### Example of suboptimality of the reference prior

Based on the discussion above, it holds that

$$\xi^{\text{Pa}} \geq \frac{e-1}{e} \left( 1 - \left( \frac{\sigma_1}{\sigma_2} \right)^{2\alpha_1} \right).$$

Recall the closed form of the  $\text{KL}_{\text{inf}}(2)$  in Lemma 4.1, which is

$$\text{KL}_{\text{inf}}(2) = \log \left( \alpha_2 \frac{\mu_1 - \sigma_2}{\mu_1} \right) + \frac{\mu_1}{\alpha_2(\mu_1 - \sigma_2)} - 1.$$

Then, let us consider the case  $\theta_1 = (1, 1.01)$  and  $\theta_2 = (10, 30)$  where  $\mu_1 = 101$  and  $\mu_2 = \frac{300}{29}$ . In this case, it holds that

$$\xi^{\text{Pa}} \gtrsim 0.626 > 0.428 \approx \frac{1}{\text{KL}_{\text{inf}}(2)},$$

which shows the suboptimality of TS based on the reference prior ( $k = 1$ ).  $\square$



#### 4.6.10 Concentration inequalities

In this section, we introduce some previously known technical results that we will use, presenting them without proof.

**Lemma 4.20** (Cramér's theorem). *Let  $X_1, \dots, X_n$  be i.i.d. random variables on  $\mathbb{R}$ . Then, for any convex set  $C \in \mathbb{R}$ ,*

$$\mathbb{P} \left[ \frac{1}{n} \sum_{i=1}^n X_i \in C \right] \leq \exp \left\{ \left( -n \inf_{z \in C} \Lambda^*(z) \right) \right\},$$

where  $\Lambda^*(z) = \sup_{\lambda \in \mathbb{R}} \lambda z - \log \mathbb{E}[e^{\lambda X_1}]$ .

**Lemma 4.21** (Bernstein's inequality). *Let  $X$  be a  $(\sigma^2, b)$ -subexponential random variable with  $\mathbb{E}[X] = \mu$  and  $\text{Var}[X] = \sigma^2$ , which satisfies*

$$\mathbb{E}[e^{\lambda X}] \leq \exp \left\{ \frac{\lambda^2 \sigma^2}{2} \right\} \quad \text{for } |\lambda| \leq \frac{1}{b}.$$

*Let  $X_i$  be independent  $(\sigma^2, b)$ -subexponential. Then, it holds that*

$$\mathbb{P} \left( \left| \frac{1}{n} \sum_{s=1}^n X_i - \mu \right| \geq t \right) \leq 2 \exp \left( -\frac{n}{2} \min \left\{ \frac{t^2}{\sigma^2}, \frac{t}{b} \right\} \right).$$

**Lemma 4.22** (Result of term (A) in Korda et al. [2013]). *When one uses the Jeffreys prior as a prior for the Pareto distribution with a known scale parameter, TS satisfies for sufficiently small  $\epsilon > 0$  that*

$$\sum_{t=1}^T \mathbb{E} [\mathbb{1}[i(t) \neq 1, \mathcal{M}_\epsilon^c(t)]] \leq \mathcal{O}(\epsilon^{-1}).$$

**Lemma 4.23** (Theorem 4.1. in Wallace [1959]). *Let  $F_n$  be the distribution function of the chi-squared distribution on  $n$  degrees of freedom. For all  $t > n$ , all  $n > 0$ , and with  $w(t) = \sqrt{t - n - n \log(t/n)}$ ,*

$$1 - F_n(t) < \frac{d_n}{2} \text{erfc} \left( \frac{w(t)}{\sqrt{2}} \right),$$

where  $d_n = \frac{\left(\frac{n}{2}\right)^{\frac{n-1}{2}} e^{-\frac{n}{2}} \sqrt{2\pi}}{\Gamma(n/2)}$  and  $\text{erfc}(\cdot)$  is the complementary error function.

#### 4.7 Conclusion

In this chapter, we extended the understanding of TS to the bandit model of the Pareto distribution that has a heavy tail and follows the power-law. While most previous research on TS has focused on one-dimensional or light-tailed distributions, we investigated the performance of TS on the Pareto distribution, which is characterized by unknown scale and shape parameters. By sequentially sampling parameters via their marginalized and conditional posterior distributions, we can realize an efficient sampling procedure. We showed that TS with the appropriate choice of priors achieves a problem-dependent optimal regret bound in such a setting for the first time. However, our investigation revealed that, as with other multiparameter models, the vanilla TS based on the reference prior and the Jeffreys prior is suboptimal. Since the reference prior is often considered the default choice in the absence of prior knowledge in multiparameter models, this suboptimality could limit its practical usage. Similar to other multiparameter

models, we found that a variant of TS proposed in Chapter 3, Thompson sampling with truncation (TS-T), can achieve the optimal performance based on the reference prior and the Jeffreys prior. This indicates the possibility of using well-known noninformative priors for other multiparameter models without having to figure out which prior is optimal, which would be a huge advantage in practice. Moreover, experimental results support the optimality of conservative priors and the effectiveness of the truncation procedure for the Jeffreys prior and the reference prior.

## Chapter 5

### Thompson Exploration

The previous chapters investigated the asymptotic optimality of Thompson sampling (TS) and how the noninformative prior affects the regret of TS for the model of the multiparameter distributions. In this chapter, we focus on the *best arm identification* (BAI) problem for the canonical single parameter exponential family (SPEF) where there exists a unique optimal arm. Although it is known that a policy designed to maximize the cumulative rewards performs poorly when exploration and evaluation phases are separated [Bubeck et al., 2011], we can still leverage the exploration part of TS as a powerful tool for solving the BAI problem.

#### 5.1 Introduction

In this section, we provide a brief introduction to the background and motivation, followed by a summary of the key contributions made in this chapter.

##### 5.1.1 Chapter background

In the BAI problem, two problem settings, the fixed-budget setting and the fixed confidence setting, have been mainly considered. As discussed in Section 2.1.2, this chapter focuses on the fixed confidence setting, where the objective is to minimize the number of trials while ensuring that the probability of misidentifying the best arm is below a fixed threshold [Even-Dar et al., 2006, Kuroki et al., 2020]. In this setting, Garivier and Kaufmann [2016] provided a tight lower bound on the expected number of trials, which is known as the sample complexity, for canonical SPEF bandit models including the Bernoulli and Gaussian distributions. This bound represents the expected number of trials required to achieve a specified level of confidence in identifying the best arm. They also proposed a policy called the Track-and-Stop (TaS) policy as an asymptotically optimal solution for the fixed confidence setting. However, it is important to note that the TaS policy involves solving computationally expensive optimization problems at each round, which may limit its practicality in real-world applications.

To address these limitations, several computationally efficient policies have been proposed, which solve the optimization problem through a single gradient ascent in the online fashion [Ménard, 2019, Wang et al., 2021a]. However, many of these policies still rely on forced exploration steps, which involve playing an arm a certain number of times to ensure the convergence of the mean estimator to its true value. Recognizing the need for a more natural exploration approach, Ménard [2019] emphasized the importance of finding methods that allow for exploration without the need for the forced exploration steps. More recently, Barrier et al. [2022] proposed a sampling policy that promotes exploration naturally by employing an upper confidence bound. However, their algorithm is specifically designed for Gaussian bandits with a known scale and exhibits slower con-

vergence of the empirical mean compared to approaches that employ the forced exploration steps. As a result, their method requires a larger number of samples in numerical experiments.

In addition to asymptotic optimality, a relaxed optimality notion, which is known as  $\beta$ -optimality, has been considered in Bayesian algorithms [Qin et al., 2017, Russo, 2016, Shang et al., 2020].  $\beta$ -optimality is commonly considered for top-two sampling rules, where the leader arm is played with a fixed probability  $\beta$  and the challenger arm is played with the probability of  $1 - \beta$ . This approach allows for different configurations of the leader and challenger arms at each round, for which more comprehensive information can be found in Jourdan et al. [2022].

Despite the advantage of not requiring the forced exploration step, top-two policies still rely on setting a hyperparameter  $\beta$  that represents the optimal proportion of the best arm, which is often unknown in practice [Qin et al., 2017, Russo, 2016, Shang et al., 2020]. Furthermore, the sample complexity bounds of these Bayesian approaches do not match the lower bound at the cost of their computational efficiency due to the relaxed nature of  $\beta$ -optimality compared to asymptotic optimality.

### 5.1.2 Chapter contribution

In this chapter, we present a simple approach that combines a heuristic method introduced by Ménard [2019] with TS using the Jeffreys prior. Our method addresses the limitations of existing approaches, which often involve solving computationally expensive optimization problems [Garivier and Kaufmann, 2016], require the forced exploration steps [Ménard, 2019, Wang et al., 2021a], or are specifically designed for Gaussian bandits [Barrier et al., 2022, Shang et al., 2020].

By combining the heuristic method and TS with the Jeffreys prior, our proposed approach eliminates the need for computationally heavy procedures and forced exploration steps. This allows for a more efficient and practical solution to the BAI problem. Furthermore, our analysis extends beyond Gaussian bandits, making our method applicable to a wider range of scenarios. It is worth noting that in the SPEF, the reference prior and the Jeffreys prior are identical [Ghosh, 2011]. However, for consistency with the previous work of Korda et al. [2013], who investigated the optimality of TS with the Jeffreys prior in the SPEF bandits, we refer to it as the Jeffreys prior throughout this chapter.

It is important to highlight that our proposed method does not achieve asymptotic optimality in all scenarios. While it exhibits near optimality, similar to the concept of  $\beta$ -optimality in Bayesian algorithms, we demonstrate that our method achieves asymptotic optimality specifically for two-armed bandit problems. Notably, our approach does not rely on additional hyperparameters (such as  $\beta$  in  $\beta$ -optimal policies), which distinguishes it from  $\beta$ -optimal policies. This unique characteristic of our method offers its own advantages and strengths compared to  $\beta$ -optimal policies.

The contributions of this chapter are summarized as follows:

- We propose a computationally efficient policy for BAI problems in SPEF bandits that eliminates the requirement of solving optimization problems, forcing explorations, and the use of additional hyperparameter  $\beta$ .
- We experimentally demonstrate the effectiveness of using TS with the Jeffreys prior as an exploration mechanism, which serves as a substitute for the forced exploration steps in BAI problems.

### 5.1.3 Chapter organization

The rest of this chapter is organized as follows. In Section 5.2, we formulate the BAI problems for the SPEF bandits and introduce the asymptotic optimality and its implications. Next, in Section 5.3, we propose a simple policy called Best Challenger with Thompson Exploration (BC-TE), which is based on a variant of the best challenger policies described in previous works [Garivier and Kaufmann, 2016, Ménard, 2019]. The sample complexity analysis of BC-TE is presented in Section 5.4, where we also compare its performance with asymptotic optimality and  $\beta$ -optimality. Furthermore, in Section 5.5, we provide simulation results that demonstrate the effectiveness of TE, showing competitive performance in terms of sample complexity and superior computation efficiency compared to other optimal policies. Finally, we provide all the detailed proofs in this chapter in Section 5.6.

## 5.2 Problem Formulation

In this section, we formulate the BAI problem for the model of SPEF and the asymptotic lower bound on the sample complexity. Then we introduce the stopping rule considered in Garivier and Kaufmann [2016].

### 5.2.1 Notation and SPEF bandits

We consider the  $K$ -armed bandit model, where each arm belongs to canonical SPEF. Here, we consider the bandit class  $\mathcal{P}$ , which is defined as follows:

$$\mathcal{P} = \left\{ (\nu_{\theta_i})_{i=1}^K : \frac{d\nu_{\theta_i}}{dP}(x) = \exp(\theta_i x - A(\theta_i)), \theta_i \in \Theta, \forall i \in [K] \right\}, \quad (5.1)$$

where  $\Theta \subset \mathbb{R}$ ,  $P$  is some reference measure on  $\mathbb{R}$ , and  $A : \Theta \rightarrow \mathbb{R}$  is a convex and twice differentiable function. For this model, we can write the expected rewards of an arm as  $\mu(\theta) = A'(\theta)$  and the KL divergence between two distributions as

$$\text{KL}(\nu_{\theta_1}, \nu_{\theta_2}) = \mu(\theta_1)(\theta_1 - \theta_2) + A(\theta_2) - A(\theta_1).$$

For simplicity, we denote the KL divergence in terms of the expected rewards by

$$d(\mu(\theta), \mu(\theta')) = \text{KL}(\nu_{\theta}, \nu_{\theta'}).$$

We denote a set of SPEF bandit models with a unique optimal arm by  $\mathcal{P}_e \subset \mathcal{P}$ . Therefore, for any  $\nu \in \mathcal{P}_e$ ,  $\arg \max_{i \in [K]} \mu(\theta_i)$  is a singleton. In this chapter, we denote the vector of the expected rewards by  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$  and assume that  $\mu(\theta_1) > \mu(\theta_2) \geq \dots \geq \mu(\theta_K)$  without loss of generality.

In the fixed-confidence setting, a policy is said to be  $\delta$  probably approximately correct ( $\delta$ -PAC) when it satisfies  $\mathbb{P}[i(\tau_\delta) \neq 1 \vee \tau_\delta = \infty] \leq \delta$ . Here,  $\tau_\delta$  is the number of trials until the sampling procedure stops for a given risk parameter  $\delta$ , and  $i(t)$  denotes the chosen arm at round  $t \in \mathbb{N}$ . Thus, the agent aims to build a  $\delta$ -PAC policy while minimizing the sample complexity  $\mathbb{E}_\nu[\tau_\delta]$ .

Here, we reproduce the notation introduced in Section 2.1.2 for readability. Recall that any  $\delta$ -PAC policy satisfies for any  $\nu \in \mathcal{P}_e$  that

$$\mathbb{E}_\nu[\tau_\delta] \geq T^*(\nu) \log \left( \frac{1}{2.4\delta} \right), \quad (5.2)$$

where

$$T^*(\nu) := \left( \sup_{\mathbf{w} \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) \right)^{-1} \quad (5.3)$$

for  $\text{Alt}(\nu) := \{\lambda \in \mathcal{P}_e : i^*(\lambda) \neq i^*(\nu)\}$ . This quantity can be written in a different way as follows.

**Lemma 5.1** (Lemma 3 in Garivier and Kaufmann [2016]). *For every  $\mathbf{w} \in \Sigma_K$ ,*

$$\inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) = \min_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu}),$$

where  $\mu_{1,i}^{\mathbf{w}} := \frac{w_1}{w_1 + w_i} \mu_1 + \frac{w_i}{w_1 + w_i} \mu_i$  is a weighted mean and

$$f_i(\mathbf{w}; \boldsymbol{\mu}) = w_1 d(\mu_1, \mu_{1,i}^{\mathbf{w}}) + w_i d(\mu_i, \mu_{1,i}^{\mathbf{w}}). \quad (5.4)$$

Hence, we can express the quantity  $T^*(\nu)$  in terms of the functions  $f_i$  and  $\boldsymbol{\mu} = \mu(\boldsymbol{\theta})$  as follows:

$$T^*(\nu) = T^*(\boldsymbol{\mu}) = \left( \sup_{\mathbf{w} \in \Sigma_K} \min_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu}) \right)^{-1}. \quad (5.5)$$

In this chapter, we will use the notations  $T^*(\nu)$  and  $T^*(\boldsymbol{\mu})$  interchangeably to represent the same quantity.

Garivier and Kaufmann [2016] also showed that the maximizer,

$$\mathbf{w}^* = \mathbf{w}^*(\boldsymbol{\mu}) := \arg \max_{\mathbf{w} \in \Sigma_K} \min_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu}),$$

indicates the optimal sampling proportion of arm draws. In other words, in order to achieve the lower bound stated in (5.2), it is necessary to adjust the proportions of arm draws at each round  $t$ , denoted by  $\mathbf{w}^t := \left( \frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t} \right)$ , to align with this optimal proportion  $\mathbf{w}^*$ . The convergence of  $\mathbf{w}^t$  towards  $\mathbf{w}^*$  is widely recognized as a crucial factor for achieving optimal performance in the BAI problem [Garivier and Kaufmann, 2016, Ménard, 2019, Wang et al., 2021a].

Based on these observations, Garivier and Kaufmann [2016] proposed the Track-and-Stop (TaS) policy that tracks the optimal proportions  $\mathbf{w}^*$  at every round and showed its asymptotic optimality. Since the true mean reward  $\boldsymbol{\mu}$  is unknown in practice, the TaS policy tracks the plug-in estimates  $\mathbf{w}^*(\hat{\boldsymbol{\mu}}(t))$ , where  $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \dots, \hat{\mu}_K(t))$  denotes the current maximum likelihood estimate of  $\boldsymbol{\mu}$  at round  $t$  for  $\hat{\mu}_i(t) = \hat{\mu}_{i, N_i(t)} = \frac{1}{N_i(t)} \sum_{s=1}^t X_{i, N_i(s)}$ . This means that the TaS policy essentially requires solving the minimax optimization problem at every round to find  $\mathbf{w}^*(\hat{\boldsymbol{\mu}}(t))$ . Although some computational burden can be alleviated by using the solution from the previous round as an initial solution, the TaS policy remains computationally expensive due to the presence of the inverse function of the KL divergence as shown in Section 5.5.

As we introduced in Section 2.1.2, a relaxed optimality notion,  $\beta$ -optimality, has been considered in Bayesian sampling rules, where the currently best arm is played with a predefined probability  $\beta \in (0, 1)$  [Jourdan et al., 2022, Qin et al., 2017, Russo, 2016, Shang et al., 2020]. Here, the counterpart of the quantity  $T^*(\boldsymbol{\mu})$  in the context of  $\beta$ -optimality is defined as follows:

$$T^\beta(\boldsymbol{\mu}) := \left( \sup_{\substack{\mathbf{w} \in \Sigma_K \\ w_1 = \beta}} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) \right)^{-1} = \left( \sup_{\substack{\mathbf{w} \in \Sigma_K \\ w_1 = \beta}} \inf_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu}) \right)^{-1}.$$

Therefore,  $\beta$ -optimal policies will satisfy  $w_1^t \rightarrow \beta$  as round  $t$  increases. However,  $\beta = 1/2$  is usually given to the algorithm in advance since  $w_1^*$  is unknown as the agent is not aware of the true expected rewards [Qin et al., 2017, Russo, 2016, Shang et al., 2020]. This means that these algorithms are not asymptotically optimal unless  $w_1^*(\boldsymbol{\mu}) = 1/2$  holds.

### 5.2.2 Stopping rule

One important question is when an agent should terminate the sampling procedure, which is usually related to a statistical test. Garivier and Kaufmann [2016] considered the generalized likelihood ratio statistic that has a closed-form expression for exponential family bandit models. Based on this statistic, they proposed Chernoff's stopping rule which is written as

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \max_{a \in [K]} \min_{b: \hat{\mu}_a(t) \geq \hat{\mu}_b(t)} t f_{a,b}(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)) > \beta(t, \delta) \right\}, \quad (5.6)$$

where  $f_{a,b}(\mathbf{w}; \boldsymbol{\mu}) := w_a d(\mu_a, \mu_{a,b}^{\mathbf{w}}) + w_b d(\mu_b, \mu_{a,b}^{\mathbf{w}})$  for  $\mu_a \geq \mu_b$  and  $\beta(t, \delta)$  is a threshold to be tuned appropriately. Therefore, several thresholds  $\beta(t, \delta)$  have been proposed [Garivier and Kaufmann, 2016, Jedra and Proutiere, 2020, Kaufmann and Koolen, 2021, Ménard, 2019]. Nevertheless, in this chapter, we simply utilize the deviational threshold

$$\beta(t, \delta) = \log \left( \frac{C t^\alpha}{\delta} \right)$$

for  $\alpha > 1$  and some constants  $C = C(\alpha, K)$  since it was shown that using Chernoff's stopping rule with this threshold ensures the  $\delta$ -PAC of any policies for the SPEF [see Garivier and Kaufmann, 2016, Proposition 12].

### 5.3 Best Challenger with Thompson Exploration

In this section, we aim to build a nearly-optimal  $\delta$ -PAC policy that does not rely on the forced exploration steps. To achieve this, we utilize TS with the Jeffreys prior as a tool, which encourages the exploration of arms in a natural manner.

For the sake of simplicity, we define a concave objective function  $g$  as

$$\begin{aligned} g : \Sigma_K \times A'(\Theta) &\rightarrow \mathbb{R}_+ \\ (\mathbf{w}; \boldsymbol{\mu}) &\mapsto \min_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu}), \end{aligned}$$

and  $g(x; \cdot) := -\infty$  for  $x \notin \Sigma_K$ . Then, (5.3) can be rewritten as

$$(T^*(\boldsymbol{\mu}))^{-1} = \sup_{\mathbf{w} \in \Sigma_K} g(\mathbf{w}; \boldsymbol{\mu}) = g(\mathbf{w}^*; \boldsymbol{\mu}).$$

As discussed in Section 5.2.1, one can achieve the asymptotic optimality by moving the empirical proportion  $\mathbf{w}^t$  closer to the optimal proportion  $\mathbf{w}^*$ . To achieve this, Garivier and Kaufmann [2016] proposed the direct tracking rule (D-tracking) that selects the arm  $i \in [K]$  that maximizes the gap  $N_i(t) - t w_i^*(\hat{\boldsymbol{\mu}}(t))$  in order to reduce this gap. However, the estimation of  $\mathbf{w}^*(\hat{\boldsymbol{\mu}}(t))$  incurs high computation costs since it requires solving the minimax optimization problem in (5.3).

Since the optimal proportion  $\mathbf{w}^*$  is a point that maximizes  $g$ , moving  $\mathbf{w}^t$  in the direction of increasing  $g$  is a reasonable idea to reduce the gap between  $\mathbf{w}^t$  and  $\mathbf{w}^*$ . As  $\mathbf{w}^*$  is a solution of a convex optimization problem, a natural approach is to apply a gradient method to update  $\mathbf{w}^t$  iteratively, which would bring  $\mathbf{w}^t$  to  $\mathbf{w}^*$  without explicitly solving complex optimization problems. Although  $g$  is not differentiable, it can be expected that playing arms to track a subgradient of  $g$  would achieve the lower bound since  $g$  is concave.<sup>1</sup>

<sup>1</sup>In the strict sense, we should use the term subgradient to minimize the convex function  $-g$  or supergradient to maximize the concave function  $g$ . However, we use the term subgradient for  $g$  since the term subgradient is more popular, and the use of  $-g$  needlessly degrades the readability.

Here, we define  $\mathbf{v}$  as a subgradient of the concave function  $g$  at the point  $(\mathbf{w}; \boldsymbol{\mu})$ , which satisfies

$$\forall \mathbf{w}' \in \Sigma_K, g(\mathbf{w}'; \boldsymbol{\mu}) \leq g(\mathbf{w}; \boldsymbol{\mu}) + \mathbf{v}^\top (\mathbf{w}' - \mathbf{w}).$$

The subdifferential  $\partial g(\mathbf{w}; \boldsymbol{\mu})$  is the set of all such subgradients. The following lemma shows that the subgradients of the objective function  $g$  are expressed as the sum of all-ones vector  $\mathbf{1}$  and convex combinations of the gradients  $\nabla_{\mathbf{w}} f_i(\mathbf{w}; \boldsymbol{\mu})$  of  $f$  with respect to  $\mathbf{w}$ . The proofs of all lemmas and theorems are given in Section 5.6.1.

**Lemma 5.2.** *The subdifferential  $\partial g$  of  $g$  with respect to  $\mathbf{w} \in \text{Int } \Sigma_K$  for given  $\boldsymbol{\mu} \in \mathcal{P}_e$  is such that*

$$\partial g(\mathbf{w}; \boldsymbol{\mu}) = \left\{ \sum_{i \in \mathcal{J}(\mathbf{w}; \boldsymbol{\mu})} \lambda_i \nabla_{\mathbf{w}} f_i(\mathbf{w}; \boldsymbol{\mu}) + r \mathbf{1} : \sum_{i \in \mathcal{J}(\mathbf{w}; \boldsymbol{\mu})} \lambda_i = 1, \lambda_i \geq 0, r \in \mathbb{R} \right\},$$

where  $\mathcal{J}(\mathbf{w}; \boldsymbol{\mu}) := \arg \min_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu})$  denotes a set of challengers,  $f_i$  is defined in (5.4), and  $\text{Int } \Sigma_K$  denotes the interior of the probability simplex.

Simply, one can consider a greedy approach that plays an arm with the maximum subgradient since our objective is to maximize the objective function  $g$ , which is

$$i(t) \in \arg \max_{i \in [K]} v_i(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)).$$

Let us define  $m(t) = \arg \max_{i \in [K]} \hat{\mu}_a(t)$  to denote the currently optimal arm at round  $t$  and  $\mathcal{J}_t = \mathcal{J}(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t))$  to denote the challengers at round  $t$ . Then, by letting  $r = 0$  in Lemma 5.2, we can obtain a subgradient  $\mathbf{v}$  satisfying

$$v_i(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)) = \begin{cases} 0 & \text{if } i \notin \{m(t)\} \cup \mathcal{J}_t, \\ \frac{1}{|\mathcal{J}_t|} \sum_{j \in \mathcal{J}_t} d(\mu_i, \mu_{i,j}^{\mathbf{w}^t}) & \text{if } i = m(t), \\ \frac{1}{|\mathcal{J}_t|} d(\mu_i, \mu_{m(t),j}^{\mathbf{w}^t}) & \text{if } i \in \mathcal{J}_t. \end{cases}$$

For notational simplicity, we denote  $\hat{\mu}_{a,b}^{\mathbf{w}^t}(t) = \frac{w_a(t)}{w_a(t)+w_b(t)} \hat{\mu}_a(t) + \frac{w_b(t)}{w_a(t)+w_b(t)} \hat{\mu}_b(t)$  by  $\hat{\mu}_{a,b}(t)$  when it is obvious in the context. Since  $\mathcal{J}_t$  is usually given as a singleton  $\{j(t)\}$ , where a challenger is

$$j(t) = \arg \min_{i \neq m(t)} f_i(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)), \quad (5.7)$$

this simple heuristic can be seen as a variant of the Best Challenger (BC) rule introduced by Garivier and Kaufmann [2016], which is equivalent to

$$i(t) = \begin{cases} m(t) & \text{if } d(\hat{\mu}_{m(t)}(t), \hat{\mu}_{m(t),j(t)}(t)) \geq d(\hat{\mu}_{j(t)}(t), \hat{\mu}_{m(t),j(t)}(t)), \\ j(t) & \text{otherwise.} \end{cases}$$

Ménard [2019] showed that this simple heuristic is computationally very efficient and shows excellent empirical performance in BAI problems.

Note that utilizing subgradients instead of solving the optimization problem at every round has been considered by Ménard [2019] where they applied the online mirror ascent method to optimize the non-smooth concave objective function  $g$ , and by Wang et al. [2021a] where they applied the Frank-Wolfe-type algorithm. Although both methods have been proven to be asymptotically optimal, they still incorporate forced exploration steps to ensure the convergence of empirical mean estimates to their true values. Hence, a key question in the BAI problem still remains:



---

**Algorithm 5** Best challenger with Thompson Exploration (BC-TE)

---

```
1: Initialization: Play every arm twice and set  $\mathbf{w}^{2K} = \frac{1}{K}$  and  $t = 2K$ .
2: while stopping criterion is satisfied do
3:   Sample  $\tilde{\mu}_a(t)$  from the posterior distribution  $\pi_{i,t}$ 
4:   Set  $m(t) = \arg \max_{i \in [K]} \hat{\mu}_a(t)$  and  $\tilde{m}(t) = \arg \max_{i \in [K]} \tilde{\mu}_i(t)$ .
5:   if  $m(t) = \tilde{m}(t)$  then
6:     Find the subgradient  $\mathbf{v}^t$  of  $g(\mathbf{w}^t, \hat{\boldsymbol{\mu}}^t)$ .
7:     Play  $i(t+1) \in \arg \max_{i \in [K]} v_i^t$  and observe the reward.
8:   else
9:     Play  $i(t+1) \in \arg \min_{i \in \{m(t), \tilde{m}(t)\}} N_i(t)$ .
10:  end if
11:  Update  $t = t + 1$ ,  $\hat{\boldsymbol{\mu}}^t$  and  $\mathbf{w}^t$ .
12: end while
```

---

Can we find a natural way to explore without forcing us to explore?

Although directly applying TS to the BAI problems leads to suboptimal performance [Bubeck et al., 2009], in this chapter, we employ TS with the Jeffreys prior, which is equivalent to the reference prior in the SPEF, as an exploration tool. To be precise, we play an arm according to the BC rule only when the empirical mean estimates and the Thompson samples agree, which is

$$i(t) = \begin{cases} \arg \max_{i \in [K]} v_i(\mathbf{w}^t; \hat{\boldsymbol{\mu}}^t) & \text{if } m(t) = \tilde{m}(t) := \arg \max_{i \in [K]} \tilde{\mu}_i(t), \\ \arg \min_{i \in \{m(t), \tilde{m}(t)\}} N_i(t) & \text{otherwise.} \end{cases}$$

As the number of plays increases, the probability of observing a sample that significantly deviates from the current empirical mean estimates exponentially decreases. In other words, when an arm is played only a few times, its Thompson sample is more likely to deviate from its mean estimate. This discrepancy between the sample and the empirical mean estimates can guide the policy to explore further. By selecting an arm with a small number of plays only when the Thompson sample and empirical estimates disagree, we can ensure the convergence of empirical mean estimates to their true values without relying on forced exploration. We formulate this result in Section 5.4. The proposed algorithm, named Best Challenger with Thompson Exploration (BC-TE), is outlined in Algorithm 5. Notice that BC-TE plays every arm twice at initialization steps to avoid an improper posterior distribution.

## 5.4 Main Theoretical Results

In this section, we show the effectiveness of Thompson exploration and prove that BC-TE is nearly optimal, similar to  $\beta$ -optimality.

### 5.4.1 Main theorems

Firstly, let us define a random variable  $T_B \in \mathbb{N}$  such that

$$\forall s \geq T_B : \sum_{i=1}^K \mathbb{1}[|\hat{\mu}_i(s) - \mu_i| \leq \epsilon] = K. \quad (5.8)$$

Therefore, the empirical mean estimate  $\hat{\boldsymbol{\mu}}(t)$  is sufficiently close to its true value  $\boldsymbol{\mu}$  for any round after  $T_B$ . The theorem below shows the expected value of  $T_B$  is finite.

**Theorem 5.3.** *Under Algorithm 5, it holds*

$$\mathbb{E}[T_B] \leq \mathcal{O}(K^2 d_\epsilon^{-2}),$$

where

$$d_\epsilon := \min_{i \in [K]} \min(d(\mu_i + \epsilon, \mu_i), d(\mu_i - \epsilon, \mu_i)). \quad (5.9)$$

Therefore, after  $T_B$  rounds, one can guess that the sampling rule will behave as expected. Then, the sample complexity of BC-TE can be upper bounded as follows.

**Theorem 5.4.** *Let  $\alpha \in [1, e/2]$  and  $r(t) = \mathcal{O}(t^\alpha)$ . Using the Chernoff's stopping rule in (5.6) with  $\beta(t, \delta) = \log(r(t)/\delta)$  under Algorithm 5,*

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \alpha \underline{T}(\nu),$$

where

$$(\underline{T}(\nu))^{-1} := \sup_{\substack{\mathbf{w} \in \Sigma_K, \\ z_{j^*(\nu)} = \gamma}} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) \quad (5.10)$$

for  $j^*(\nu) = \arg \max_{i \neq 1} \mu_i$ ,  $z_i := \frac{w_i}{w_1 + w_i}$ , and

$$d(\mu_1, (1 - \gamma)\mu_1 + \gamma\mu_{j^*(\nu)}) = d(\mu_{j^*(\nu)}, (1 - \gamma)\mu_1 + \gamma\mu_{j^*(\nu)}). \quad (5.11)$$

Recall that in this chapter, we assume  $j^*(\nu) = 2$  for notational simplicity. From the definition of  $T^*(\nu)$  in (5.3), which satisfies for  $\gamma$  in (5.11)

$$\begin{aligned} (T^*(\nu))^{-1} &= \sup_{\mathbf{w} \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) \\ &\geq \sup_{\substack{\mathbf{w} \in \Sigma_K, \\ z_{j^*(\nu)} = \gamma}} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) = (\underline{T}(\nu))^{-1}. \end{aligned}$$

Therefore, one can see the suboptimality of BC-TE from

$$\underline{T}(\nu) \geq T^*(\nu),$$

which indicates that BC-TE may not always be optimal, as it only achieves optimality when the condition  $z_2 = \frac{w_2^*}{w_1^* + w_2^*}$  is true. This observation is akin to the result for  $\beta$ -optimality. In the following sections, we will further explore the implications of this quantity  $\underline{T}(\nu)$ .

#### 5.4.2 Comparison with $\beta$ -optimality

Here, we provide an interpretation of the upper bound on the sample complexity of BC-TE by comparing it with  $\beta$ -optimality.

Recall that a policy is  $\beta$ -optimal if it satisfies

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq T^\beta(\nu) = \left( \sup_{\substack{\mathbf{w} \in \Sigma_K, \\ w_{i^*(\nu)} = \beta}} \inf_{\lambda \in \text{Alt}(\nu)} \left( \sum_{i=1}^K w_i \text{KL}(\nu_i; \lambda_i) \right) \right)^{-1}.$$

It is important to note that  $\beta$ -optimality is achieved when the allocation of the optimal arm is  $\beta$ . On the other hand,  $\underline{T}(\nu)$  considers the scenario where the allocation of the

second-best arm is  $\gamma$ , as defined in (5.11). Therefore, both notions are more relaxed compared to asymptotic optimality, and it is not possible to determine definitively which one is better in general.

However, it is important to note that our policy does not require prior knowledge of  $\gamma$ , differently from existing  $\beta$ -optimal policies that take  $\beta$  as an input to the algorithm [Jourdan and Degenne, 2022, Jourdan et al., 2022, Russo, 2016, Shang et al., 2020]. Therefore, unless there is prior knowledge of the specific value of  $\beta$ , it would be advantageous to use BC-TE instead of  $\beta$ -optimal policies. BC-TE naturally addresses the BAI problem in a nearly optimal manner without the need for a hyperparameter  $\beta$  or forcing exploration.

### 5.4.3 Comparison with asymptotic optimality

The natural question is the relationship between  $T^*(\nu)$  and  $\underline{T}(\nu)$ . Differently from the  $\beta$ -optimality where  $\beta$  does not depend on the bandit instance, our near-optimality is a problem-dependent notion. Here, we provide a rough comparison with the asymptotic optimality.

First, let us introduce a function that enables us to derive a more explicit formula for  $w^*(\mu)$ , for any  $i \neq 1$ ,

$$k_i(x; \mu) = d\left(\mu_1, \frac{1}{1+x}\mu_1 + \frac{x}{1+x}\mu_i\right) + xd\left(\mu_i, \frac{1}{1+x}\mu_1 + \frac{x}{1+x}\mu_i\right).$$

As demonstrated in Garivier and Kaufmann [2016], this function is a strictly increasing bijective mapping from  $[0, \infty)$  onto  $[0, d(\mu_1, \mu_a))$ . Therefore, one can define  $l_i$  as the inverse function of  $k_i$  for any  $i \neq 1$  and  $l_1$  as a constant function, which is

$$\begin{aligned} k_i^{-1} &= l_i : [0, d(\mu_1, \mu_i)) \mapsto [0, \infty) \\ l_1 &: [0, d(\mu_1, \mu_i)) \mapsto 1. \end{aligned} \tag{5.12}$$

Then, Garivier and Kaufmann [2016] provided the following characterization of  $w^*(\mu)$ .

**Lemma 5.5** (Theorem 5 in Garivier and Kaufmann [2016]). *For every  $i \in [K]$ ,*

$$w_i^*(\mu) = \frac{l_i(y^*)}{\sum_{a=1}^K l_a(y^*)},$$

where  $y^*$  is the unique solution of the equation  $F_\mu(y) = 1$ , and where

$$F_\mu : y \mapsto \sum_{i=2}^K \frac{d\left(\mu_1, \frac{\mu_1 + l_i(y)\mu_i}{1 + l_i(y)}\right)}{d\left(\mu_i, \frac{\mu_1 + l_i(y)\mu_i}{1 + l_i(y)}\right)}$$

is a continuous, increasing function on  $[0, d(\mu_1, \mu_2))$  such that  $F_\mu(0) = 0$  and  $F_\mu(y) = \infty$  when  $y \rightarrow d(\mu_1, \mu_2)$ .

However, to derive a more explicit formula for the maximizer of (5.10), we require another function for any  $i \neq 1$

$$h_i(z; \mu) = (1-z)d(\mu_1, (1-z)\mu_1 + z\mu_i) + zd(\mu_i, (1-z)\mu_1 + z\mu_i),$$

whose domain is  $[0, 1]$ . The derivative of this function is

$$h'_i(z; \mu) = d(\mu_i, (1-z)\mu_1 + z\mu_i) - d(\mu_1, (1-z)\mu_1 + z\mu_i).$$

Thus,  $h_i(z; \boldsymbol{\mu})$  is a concave function with  $h_i(0; \boldsymbol{\mu}) = 0$  and  $h_i(1, \boldsymbol{\mu}) = 0$ . It reaches its maximum at

$$z_i^*(\boldsymbol{\mu}) : d(\mu_i, (1 - z_i^*)\mu_1 + z_i^*\mu_i) = d(\mu_1, (1 - z_i^*)\mu_1 + z_i^*\mu_i). \quad (5.13)$$

Therefore, one can see that  $\gamma = z_2^*$ .

From the definitions of  $f_i$ ,  $k_i$ , and  $h_i$ , one can find the following relationship

$$f_i(\mathbf{w}; \boldsymbol{\mu}) = w_1 k_i\left(\frac{w_i}{w_1}; \boldsymbol{\mu}\right) = (w_1 + w_i) h_i\left(\frac{w_i}{w_1 + w_i}; \boldsymbol{\mu}\right). \quad (5.14)$$

For  $z_i = \frac{w_i}{w_1 + w_i}$ , the equality between  $h_i$  and  $k_i$  can be written as

$$h_i(z_i; \boldsymbol{\mu}) = (1 - z_i) k_i\left(\frac{z_i}{1 - z_i}; \boldsymbol{\mu}\right).$$

We further define the problem-dependent constant  $\underline{z}_i \in [0, 1]$  satisfying

$$\underline{z}_i : k_i\left(\frac{\underline{z}_i}{1 - \underline{z}_i}; \boldsymbol{\mu}\right) = k_2\left(\frac{z_2^*}{1 - z_2^*}; \boldsymbol{\mu}\right), \quad (5.15)$$

for  $i \neq 1$  and  $\underline{z}_1 = \frac{1}{2}$ . Here, we have  $\underline{z}_2 = z_2^*$  and  $\underline{z}_i \leq z_2^*$  since  $k_i$  is strictly increasing and  $k_i(x; \boldsymbol{\mu}) \leq k_j(x; \boldsymbol{\mu})$  holds for any  $x \in \mathbb{R}_+$  if  $\mu_i \leq \mu_j$  [see Garivier and Kaufmann, 2016, Appendix A.3.]. Based on  $\underline{z}_i$ , we can define the normalized proportion  $\underline{w} \in \Sigma_K$  by

$$\underline{w}_i(\boldsymbol{\mu}) = \frac{\frac{\underline{z}_i}{1 - \underline{z}_i}}{\sum_{i=1}^K \frac{\underline{z}_i}{1 - \underline{z}_i}} = \frac{l_i(\underline{y})}{\sum_{i=1}^K l_i(\underline{y})}, \quad (5.16)$$

where  $\underline{y} = k_i\left(\frac{\underline{z}_i}{1 - \underline{z}_i}; \boldsymbol{\mu}\right)$  for any  $i \neq 1$ . Therefore, Theorem 5.4 implies that the empirical proportion of arm plays of BC-TE will converge to  $\underline{w}$ , which is equivalent to

$$g(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)) \rightarrow g(\underline{w}; \boldsymbol{\mu}).$$

Here, one can see that  $F_{\boldsymbol{\mu}}(\underline{y}) \geq 1$  since

$$\frac{d\left(\mu_1, \frac{\mu_1 + \frac{\underline{z}_2}{1 - \underline{z}_2} \mu_2}{1 + \frac{\underline{z}_2}{1 - \underline{z}_2}}\right)}{d\left(\mu_2, \frac{\mu_1 + \frac{\underline{z}_2}{1 - \underline{z}_2} \mu_2}{1 + \frac{\underline{z}_2}{1 - \underline{z}_2}}\right)} = \frac{d(\mu_1, (1 - \underline{z}_2)\mu_1 + \underline{z}_2\mu_2)}{d(\mu_2, (1 - \underline{z}_2)\mu_1 + \underline{z}_2\mu_2)} = 1$$

holds from the definition of  $\underline{z}_2 = z_2^*$  in (5.13), which directly implies that  $\underline{y} \geq y^*$ . However, it is important to note that from  $\underline{z}_i \leq z_i^*$ , it always hold that for any  $i \neq 1$

$$\frac{d(\mu_1, (1 - \underline{z}_i)\mu_1 + \underline{z}_i\mu_i)}{d(\mu_i, (1 - \underline{z}_i)\mu_1 + \underline{z}_i\mu_i)} \leq \frac{d(\mu_1, (1 - z_i^*)\mu_1 + z_i^*\mu_i)}{d(\mu_i, (1 - z_i^*)\mu_1 + z_i^*\mu_i)} = 1.$$

This implies that

$$1 \leq F_{\boldsymbol{\mu}}(\underline{y}) \leq K - 1,$$

where the right equality holds only when  $\mu_2 = \mu_3 = \dots = \mu_K$ . Here, it is important to note that the left equality is always valid for two-armed bandit problems. In other words, BC-TE is *asymptotically optimal* in the context of two-armed bandit problems.

**Example. Gaussian bandits**

In Gaussian bandits, the KL-divergence takes a simple form:  $d(\mu, \mu') = \frac{(\mu - \mu')^2}{2\sigma^2}$ , where for any  $i \neq 1$

$$k_i(x; \boldsymbol{\mu}) = \left( \frac{x}{1+x} \right)^2 \frac{\Delta_i^2}{2\sigma^2} + \frac{x}{(1+x)^2} \frac{\Delta_i^2}{2\sigma^2} = \frac{x}{1+x} \frac{\Delta_i^2}{2\sigma^2}$$

$$h_i(z; \boldsymbol{\mu}) = z(1-z) \frac{\Delta_i^2}{2\sigma^2}.$$

This simple formulation allows us to derive a more explicit comparison with asymptotic optimality.

Firstly, the maximizers of  $h_i, z_i^*$  in (5.13) is written as

$$\frac{(\mu_1 - \mu_i)^2}{2\sigma^2} (1 - z_i^*)^2 = \frac{(\mu_1 - \mu_i)^2}{2\sigma^2} (z_i^*)^2,$$

which implies that  $z_i^* = 1/2$  for any  $i \neq 1$ . Then, for any  $i \neq 1$ , from the definition of  $\underline{z}_i$  in (5.15), it holds

$$k_2(1; \boldsymbol{\mu}) = \frac{\Delta_2^2}{4\sigma^2} = k_i \left( \frac{\underline{z}_i}{1 - \underline{z}_i}; \boldsymbol{\mu} \right)$$

$$= \frac{\Delta_i^2}{2\sigma^2} \underline{z}_i,$$

which implies  $\underline{z}_i = \frac{\Delta_2^2}{2\Delta_i^2}$  for  $i \neq 1$ . Therefore,

$$\underline{w}_i = \frac{\frac{\Delta_2^2}{2\Delta_i^2 - \Delta_2^2}}{\sum_{a=1}^K \frac{\Delta_a^2}{2\Delta_a^2 - \Delta_2^2}}.$$

Then the objective function is written as

$$g(\underline{w}; \boldsymbol{\mu}) = \underline{w}_1 k_1 \left( \frac{\underline{z}_1}{1 - \underline{z}_1}; \boldsymbol{\mu} \right) = \frac{1}{\sum_{a=1}^K \frac{\Delta_a^2}{2\Delta_a^2 - \Delta_2^2}} \frac{\Delta_2^2}{4\sigma^2},$$

which implies that

$$\underline{T}(\boldsymbol{\mu}) = \sum_{i=1}^K \frac{4\sigma^2}{\Delta_i^2 + (\Delta_i^2 - \Delta_2^2)}.$$

Next, Garivier and Kaufmann [2016] showed the following inequalities hold for the Gaussian bandits if  $\Delta_1 = \Delta_2$  and  $\Delta_i = \mu_1 - \mu_i$

$$\sum_{i=1}^K \frac{2\sigma^2}{\Delta_i^2} \leq T^*(\boldsymbol{\mu}) \leq 2 \sum_{i=1}^K \frac{2\sigma^2}{\Delta_i^2},$$

which directly implies that

$$T^*(\boldsymbol{\mu}) \leq \underline{T}(\boldsymbol{\mu}) \leq 2T^*(\boldsymbol{\mu}),$$

where the left equality holds when  $w_1^*(\boldsymbol{\mu}) = w_2^*(\boldsymbol{\mu})$  and the right equality holds only when  $\mu_2 = \dots = \mu_K$ . Notice that this is the same result as the  $\beta$ -optimal policy when  $\beta$  is given as  $1/2$  [Russo, 2016].

## 5.5 Simulation Results

In this section, we present numerical results to demonstrate the performance of BC-TE.

**Competing algorithms:** We compare the performance of BC-TE with other policies, where  $\dagger$ ,  $\ddagger$  denotes that the policy is implemented by Koolen [2019] or by Wang et al. [2021a], respectively, and  $\diamond$  denotes that the policy requires the forced exploration:

- Track-and-Stop $\dagger, \diamond$  (TaS): an asymptotically optimal policy that solves the optimization problem in (5.3) at every round, which is very costly [Garivier and Kaufmann, 2016].
- Lazy Mirror Ascent $\dagger, \diamond$  (LMA): a computationally efficient and asymptotically optimal algorithm that performs a single gradient ascent in an online fashion [Ménard, 2019].
- AdaHedge vs Best Response $\dagger$  (AHBR): an asymptotically optimal algorithm that solves the optimization problem as an unknown game [Degenne et al., 2019a].
- Optimistic TaS $\ddagger$  (O-C): The optimistic TaS policies with C-tracking proposed by Degenne et al. [2019a], which is known to be highly computationally expensive.
- Frank-Wolfe Sampling $\ddagger, \diamond$  (FWS): an asymptotically optimal algorithm that just relies on a single iteration FW algorithm instead of solving the optimization problems in (5.3) at every round [Wang et al., 2021a].
- Round Robin (RR): a simple baseline that samples arms in a round-robin manner.
- Top-Two Transportation Cost (T3C): a computationally efficient asymptotically  $\beta$ -optimal top-two algorithm based on TS [Shang et al., 2020]. Notice that its  $\beta$ -optimality is proven only for the Gaussian models with a known scale.

While there exist two versions of the TaS policy, we focus on the TaS with D-tracking (T-D) in our experiments. T-D directly tracks the optimal proportion of arm draws at each round ( $N(t) \rightsquigarrow t\mathbf{w}^*(\hat{\boldsymbol{\mu}}(t))$ ), and it has been found to outperform the version with C-tracking in experiments, which tracks the cumulative optimal proportions ( $N(t) \rightsquigarrow \sum_{s \leq t} \mathbf{w}^*(\hat{\boldsymbol{\mu}}(s))$ ).

Furthermore, we propose a modified version of FWS, called FWS-TE, where we replace the forced exploration step in FWS with our Thompson exploration step. This adaptation is based on our conjecture that if we integrate TE into a policy that selects an arm to increase the objective function at every round, Theorem 5.3 can be derived.

**Stopping rule:** Following the experiments in the previous researches [Degenne et al., 2019a, Garivier and Kaufmann, 2016, Ménard, 2019, Wang et al., 2021a], we considered the same threshold  $\beta(t, \delta) = \log((\log(t) + 1)/\delta)$ .

**General setup:** In this section, we provide the empirical sample complexities of various policies for a range of risk levels  $\delta \in \{0.2, 0.1, 0.01, 0.001\}$  averaged over 3,000 independent runs. Following Degenne et al. [2019a], we consider the practical version of the lower bound (PLB), which refers to the first round where  $tg(\mathbf{w}^*; \boldsymbol{\mu}) \geq \beta(t, \delta)$  is satisfied. Hence, this practical lower bound indicates the earliest round where the generalized likelihood ratio statistic approximately crosses the threshold, and is defined as round  $s$  where  $s = \beta(s, \delta)T^*(\boldsymbol{\mu})$  holds. Recall that the asymptotic lower bound is given as  $T^*(\boldsymbol{\mu}) \log(\frac{1}{2.4\delta})$  according to (5.2).

Table 5.1: Sample complexity for 5-armed Bernoulli bandit instance with means  $\mu^B = (0.3, 0.21, 0.2, 0.19, 0.18)$  over 3,000 independent runs. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a].

$\delta$	BC-TE	FWS-TE	FWS	LMA	T-D	O-C	AHBR	T3C	RR	PLB	LB
0.2	1065	1077	1176	1415	1107	1545	1615	1115	1977	1208	272
0.1	1288	1326	1373	1668	1337	1818	1859	1372	2326	1442	574
0.01	2064	2102	2125	2509	2066	2706	2675	2180	3460	2211	1471
0.001	2849	2870	2880	3362	2823	3584	3469	3011	4555	2974	2252

Table 5.2: Sample complexity for 4-armed Gaussian bandit instance with means  $\mu_4^G = (1.0, 0.85, 0.8, 0.7)$  with unit scale over 3,000 independent runs. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a].

$\delta$	BC-TE	FWS-TE	FWS	LMA	T-D	O-C	AHBR	T3C	RR	PLB	LB
0.2	1415	1435	1499	1799	1472	1837	1959	1482	2555	1683	374
0.1	1759	1772	1829	2153	1806	2235	2339	1833	3078	2004	791
0.01	2895	2887	2890	3300	2835	3501	3524	2947	4730	3062	2026
0.001	3987	3967	3922	4445	3908	4732	4657	4042	6349	4112	3101

**Bernoulli models** In the first experiment, we consider the 5-armed Bernoulli bandit instance  $\nu_5^B$  where  $\mu_5^B = (0.3, 0.21, 0.2, 0.19, 0.18)$  and  $w^*(\mu_5^B) = (0.43, 0.25, 0.18, 0.13, 0.10)$ , which follows the experiment considered in previous researches [Garivier and Kaufmann, 2016, Ménard, 2019, Wang et al., 2021a].

**Gaussian models** In the second experiment, we consider the 4-armed Gaussian bandit instance  $\nu_4^G$  with unit scale  $\sigma = 1$  where  $\mu_4^G = (1.0, 0.85, 0.8, 0.7)$  and  $w^*(\mu_4^G) = (0.41, 0.38, 0.15, 0.06)$ , which was studied in Wang et al. [2021a].

**Results** The overall results for each model are presented in Tables 5.1 for Bernoulli bandits and 5.2 for Gaussian bandits, respectively. Although our proposed method BC-TE does not achieve the asymptotic optimality in general, it exhibits a better empirical performance than other optimal policies across all risk parameters, especially when large  $\delta$  is considered. Note that T3C does not have any theoretical guarantee for the Bernoulli bandit cases, while our proposed method, BC-TE, has been demonstrated to be near-optimal for the SPEF.

Interestingly, Figures 5.1 and 5.2 show that both BC-TE and FWS-TE consistently outperform other optimal policies, demonstrating the practical effectiveness of TE as an alternative to the forced exploration steps. Furthermore, we observe that BC-TE is more computationally efficient than other asymptotically optimal policies, and FWS-TE outperforms the original FWS in terms of efficiency, as demonstrated in Table 5.3.

Table 5.3: Average time of one step of various policies.

Policy	Instance	BC-TE	FWS-TE	FWS	LMA	T-D	O-C	AHBR	T3C	RR
Time (relative)	$\mu_5^B$	1	35.53	40.13	1.743	43.52	448.1	2.695	0.843	0.328
	$\mu_4^G$	1	80.77	96.30	3.588	582.3	4533	3.935	0.71	0.425

## 5.6 Proofs of Theoretical Results

In this section, we provide detailed proofs for the results in Section 5.4.

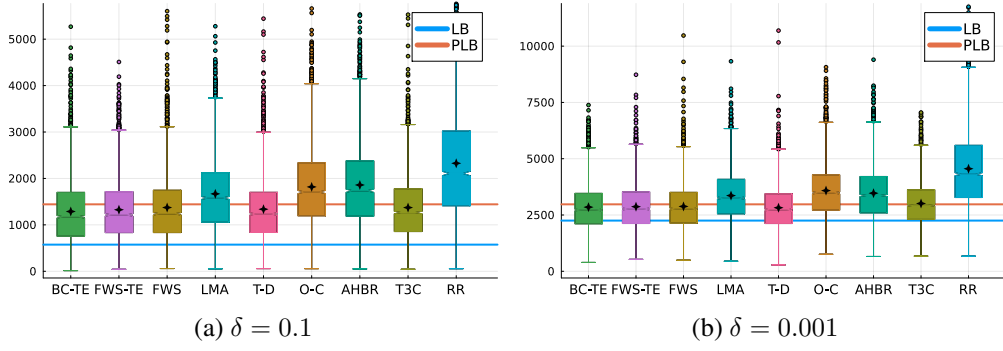


Figure 5.1: Stopping times of various policies for 5-armed Bernoulli bandit instance with means  $\mu_5^B = (0.3, 0.21, 0.2, 0.19, 0.18)$  and different maximal risk over 3,000 independent runs. The black star denotes the mean of stopping times. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a].

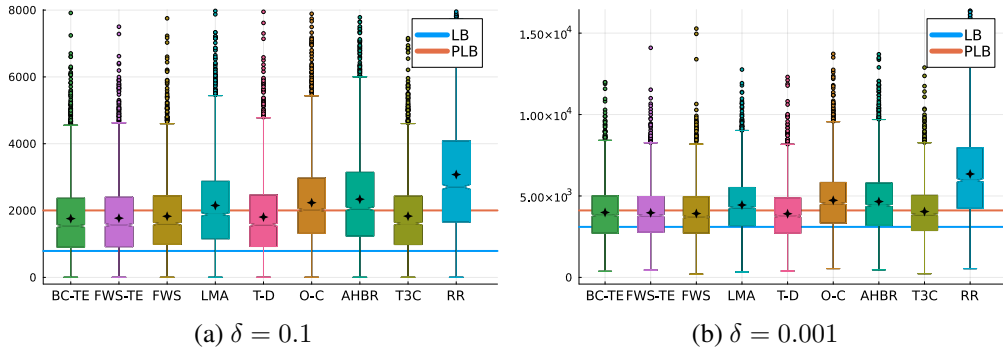


Figure 5.2: Stopping times of various policies for 4-armed Gaussian bandit instance with means  $\mu_4^G = (1.0, 0.85, 0.8, 0.7)$  and unit scale and different maximal risk over 3,000 independent runs. The black star denotes the mean of stopping times. LB denotes the lower bound given in (5.2), and PLB denotes the practical version of LB considered in Degenne et al. [2019a].

Before beginning the proof, we first define good events on estimates  $\hat{\mu}_i(t)$  and Thompson samples  $\tilde{\mu}_i(t)$  for any  $\epsilon > 0$ ,

$$\begin{aligned}\mathcal{A}_i(t) &= \mathcal{A}_{i,\epsilon}(t) := \begin{cases} \{\hat{\mu}_1(t) \geq \mu_1 - \epsilon\}, & \text{if } i = 1, \\ \{\hat{\mu}_i(t) \leq \mu_i + \epsilon\}, & \text{otherwise,} \end{cases} \\ \mathcal{B}_i(t) &= \mathcal{B}_{i,\epsilon}(t) := \{|\hat{\mu}_i(t) - \mu_i| \leq \epsilon\}, \\ \tilde{\mathcal{B}}_i(t) &= \tilde{\mathcal{B}}_{i,\epsilon}(t) := \{|\tilde{\mu}_i(t) - \mu_i| \leq \epsilon\}, \\ \mathcal{M}(t) &:= \{m(t) = \tilde{m}(t)\},\end{aligned}$$

Note that for all  $i \in [K]$  and  $t \in \mathbb{N}$ ,  $\mathcal{B}_i(t) \subset \mathcal{A}_i(t)$  holds.

Next, let us define another random variables  $D_1 = D_{1,\epsilon} := \max_{i \neq 1} D_{i,\epsilon}$  where

$$D_i = D_{i,\epsilon} := \sup_{t \geq 2K+1} \mathbb{1}[\mathcal{B}_{i,\epsilon}^c(t)] N_i(t) d(\hat{\mu}_i(t), \hat{\mu}_1(t))$$

denotes the supremum of  $N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_1(t))$  when  $\mathcal{B}_{i,\epsilon}^c(t)$  occurs. In other words,

$$\{N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_1(t)) \geq D_{i,\epsilon}\} \implies \{\mathbb{1}[\mathcal{B}_{i,\epsilon}(t)] = 1\}.$$



We further define  $\underline{d}_1 = d(\mu_1 - \epsilon, \mu_2 + \epsilon)$  and for  $i \neq 1$

$$\underline{d}_i = \min_{\substack{\mu \in [\mu'_i, \mu'_1], \\ \mu'_i \leq \mu_i + \epsilon, \mu'_1 \geq \mu_1 - \epsilon, \\ d(\mu'_i, \mu) \geq d(\mu'_1, \mu)}} d(\mu'_i, \mu). \quad (5.17)$$

### 5.6.1 Proof of Lemma 5.2

Here, we derive the subdifferential of the objective function  $g$ .

*Proof.* By abuse of notation, we define a characteristic function  $I_{\Sigma_K} : \mathbb{R}^K \rightarrow \mathbb{R}$ ,

$$I_{\Sigma_K}(x) = \begin{cases} 0, & \text{if } x \in \Sigma_K \\ -\infty, & \text{if } x \notin \Sigma_K. \end{cases}$$

Then, the problem in (5.5) can be written as

$$\sup_{\mathbf{w} \in \Sigma_K} \min_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu}) = \max_{\mathbf{w} \in \mathbb{R}^K} \left\{ \min_{i \neq 1} f_i(\mathbf{w}) + I_{\Sigma_K}(\mathbf{w}) \right\}. \quad (5.18)$$

Then, the set of differential of (5.18) is

$$\partial \left( \min_{a \neq 1} f_a(\mathbf{w}) + I_{\Sigma_K}(\mathbf{w}) \right) = \left\{ q + r : q \in \partial \min_{i \neq a} f_a(\mathbf{w}), r \in \partial I_{\Sigma_K}(\mathbf{w}) \right\}.$$

Let  $\partial I_{\Sigma_K}(\mathbf{w})$  denote the set of subgradient  $\mathbf{v}$  of  $I_{\Sigma_K}$  at point  $(\mathbf{w}; \boldsymbol{\mu})$ . Then,  $\partial I_{\Sigma_K}(\mathbf{w})$  is written as

$$\partial I_{\Sigma_K}(\mathbf{w}) = \{ \mathbf{v} \in \mathbb{R}^K : \forall \mathbf{x} \in \mathbb{R}^K, I_{\Sigma_K}(\mathbf{x}) \leq I_{\Sigma_K}(\mathbf{w}) + \mathbf{v}^\top (\mathbf{x} - \mathbf{w}) \} \quad (5.19)$$

From the definition of  $I_{\Sigma_K}$ , if  $\mathbf{x} \notin \Sigma_K$ , the inequality constraint in (5.19) always holds for any  $\mathbf{v} \in \mathbb{R}^K$ . Thus, it suffices to show that

$$\begin{aligned} \partial I_{\Sigma_K}(\mathbf{w}) &= \{ \mathbf{v} \in \mathbb{R}^K : \forall \mathbf{x} \in \Sigma_K, I_{\Sigma_K}(\mathbf{x}) \leq I_{\Sigma_K}(\mathbf{w}) + \mathbf{v}^\top (\mathbf{x} - \mathbf{w}) \} \\ &= \{ r \mathbf{1} : r \in \mathbb{R} \} \end{aligned} \quad (5.20)$$

holds, which implies that all subgradients  $\mathbf{v}$  can be written as a multiple of the all-one vector  $\mathbf{1} = [1, \dots, 1] \in \mathbb{R}^K$ .

$$(1) \{ r \mathbf{1} : r \in \mathbb{R} \} \subset \partial I_{\Sigma_K}(\mathbf{w})$$

Note that  $\mathbf{0} \in \partial I_{\Sigma_K}(\mathbf{w})$ , which implies  $\partial I_{\Sigma_K}(\mathbf{w}) \neq \emptyset$ . Since  $\mathbf{x} \in \Sigma_K$ ,  $\mathbf{v} \in \partial I_{\Sigma_K}(\mathbf{w})$  satisfies  $0 \leq \mathbf{v}^\top (\mathbf{x} - \mathbf{w})$  for all  $\mathbf{x} \in \Sigma_K$ . One can see that  $\{ r \mathbf{1} : r \in \mathbb{R} \} \subset \partial I_{\Sigma_K}(\mathbf{w})$  for  $\mathbf{w} \in \Sigma_K$ .

$$(2) \{ r \mathbf{1} : r \in \mathbb{R} \} \supset \partial I_{\Sigma_K}(\mathbf{w})$$

Then, we need to show the equality in (5.20) for  $\mathbf{w} \in \text{Int } \Sigma_K$ . At first, let assume  $K \geq 2$  and  $\mathbf{v} = r \mathbf{1} + \sum_{i=1}^K a_i e_i$ , where  $e_i$  is a standard basis for  $\mathbb{R}^K$  and  $a_i \in \mathbb{R}$ . Then,  $\forall \mathbf{x} \in \Sigma_K$ ,

$$0 \leq \sum_{i=1}^K a_i (x_i - w_i) \quad (5.21)$$

holds. We will prove the equality in (5.20) by contradiction, i.e., we assume that there exist  $i \neq j \in [K]$  such that  $a_i \neq a_j$ . From the definition of  $\text{Int}\Sigma_K$ , we can take a positive constant  $\epsilon \in \mathbb{R}_+$  satisfying  $0 < \epsilon < \min(\min_i w_i, 1 - \max_i(w_i))$ .<sup>2</sup>

Define two  $K$  dimensional vectors as

$$\mathbf{x}^1 = (x_i)_{i=1}^K = \begin{cases} w_i, & \text{if } i \in [K] \setminus \{i_1, i_2\}, \\ w_i + \epsilon, & \text{if } i = i_1, \\ w_i - \epsilon, & \text{if } i = i_2, \end{cases}$$

and

$$\mathbf{x}^2 = (x_i)_{i=1}^K = \begin{cases} w_i, & \text{if } i \in [K] \setminus \{i_1, i_2\}, \\ w_i - \epsilon, & \text{if } i = i_1, \\ w_i + \epsilon, & \text{if } i = i_2, \end{cases}$$

where  $i_1 \neq i_2 \in [K]$ . Then, both  $\mathbf{x}^1$  and  $\mathbf{x}^2$  are in  $\Sigma_K$ . From (5.21), this implies that two inequalities

$$0 \leq \epsilon(a_{i_1} - a_{i_2}) \text{ and } 0 \leq -\epsilon(a_{i_1} - a_{i_2})$$

hold at the same time. Thus,  $a_{i_1} = a_{i_2}$  should hold. However, we can make these kinds of vectors for every pair of bases, which means that  $\nexists i \neq j \in [K]$  such that  $a_i \neq a_j$ . This is a contradiction, and thus (5.20) holds.

### (3) Conclusion

Consequently, it holds  $\forall \mathbf{w} \in \text{Int}\Sigma_K$  that

$$\begin{aligned} \partial g &= \left\{ q + r\mathbf{1} : q \in \mathbf{Co} \bigcup \{ \partial f_i(\mathbf{w}; \boldsymbol{\mu}) : f_i(\mathbf{w}; \boldsymbol{\mu}) = g(\mathbf{w}; \boldsymbol{\mu}) \}, r \in \mathbb{R} \right\} \\ &= \left\{ q + r\mathbf{1} : q \in \mathbf{Co} \bigcup \{ \nabla_{\mathbf{w}} f_i(\mathbf{w}; \boldsymbol{\mu}) : f_i(\mathbf{w}) = g(\mathbf{w}) \}, r \in \mathbb{R} \right\}, \end{aligned}$$

where  $\mathbf{Co} \bigcup \{ \nabla_{\mathbf{w}} f_i(\mathbf{w}; \boldsymbol{\mu}) : f_i(\mathbf{w}; \boldsymbol{\mu}) = g(\mathbf{w}; \boldsymbol{\mu}) \}$  is the convex hull of the union of superdifferentials of all active function at  $\mathbf{w}$ . Let us define the set

$$\mathcal{J}(\mathbf{w}; \boldsymbol{\mu}) := \arg \min_{i \neq 1} f_i(\mathbf{w}; \boldsymbol{\mu}) = \{i \in [K] : f_i = g\},$$

which concludes the proof. □

#### 5.6.2 Proof of Theorem 5.3: Convergence of estimates

We begin the proof of Theorem 5.3 by introducing two lemmas that show a sufficient condition to occur  $\mathcal{B}_i(t)$  for  $i = 1$  and  $i \neq 1$ , respectively.

**Lemma 5.6.** *For any constant  $M > 0$ , assume that*

$$\{m(t) = 1, j(t) = j, i(t) = j, \mathcal{A}_1(t), \mathcal{B}_j(t), \mathcal{M}(t), N_j(t) > \max\{M, D_1/d_j\}\}$$

*occurred for some  $t$ . Then, for all  $t' \geq t$ , we have  $\mathbb{1}[\mathcal{B}_1(t')] = 1$  and*

$$N_1(t) \geq \frac{\max\{d_j M, D_1\}}{d(\mu_1 + \epsilon, \mu_j - \epsilon)}.$$

---

<sup>2</sup>Note that such  $\epsilon$  always exists by Archimedean property if  $\mathbf{w}$  is in the interior of the probability simplex, i.e.,  $\forall i \in [K], w_i \neq 0, 1$ .

**Lemma 5.7.** For any constant  $M > 0$ , assume that

$$\left\{ m(t) = 1, i(t) = 1, \mathcal{A}_{j(t)}(t), \mathcal{B}_1(t), \mathcal{M}(t), N_1(t) > \max \left\{ M, \max_{i \neq 1} \frac{D_i}{\underline{d}_i} \right\} \right\}$$

occurred for some  $t$ . Then, for all  $i \neq 1$  and  $t' \geq t$ , we have  $\mathbb{1}[\mathcal{B}_i(t')] = 1$  and

$$N_i(t) \geq \frac{\max\{\underline{d}_i M, D_i\}}{d(\mu_1 + \epsilon, \mu_i - \epsilon)}.$$

Therefore, if both events in Lemmas 5.6 and 5.7 occurred until rounds  $T$ , only  $\{\mathcal{B}_i(t)\}$  occurs for all  $i \in [K]$  and  $t \geq T$ . The proofs of these lemmas are postponed to Section 5.6.3.

*Proof of Theorem 5.3.* Firstly, let us define another random variable  $T_C \leq T_B$  such that

$$\forall s \geq T_C : \mathbb{1}[\mathcal{B}_1(s)] = 1,$$

which implies that the mean estimate of the optimal arm is close to its true value after  $T_C$  rounds. Let  $D = \max \left\{ M, \frac{D_1}{\min_{a \in [K]} \underline{d}_a} \right\}$  for some positive constant  $M$  specified later and  $T_M = \max(KD, T_C)$ . Let us consider a subset of rounds with any fixed  $T > T_M$

$$\begin{aligned} S_1(T) &:= \{s \in [T_M, T] \cap \mathbb{N} : m(s) = 1, i(s) = j(s), \mathcal{B}_1(s), \mathcal{B}_{j(s)}(s), \mathcal{M}(s)\} \\ &= \{T_{S_1} =: s_{S_1,1}, s_{S_1,2}, \dots, s_{S_1,|S_1(T)|}\} \\ S_2(T) &:= \{s \in [T_M, T] \cap \mathbb{N} : m(s) = 1, i(s) = 1, \mathcal{A}_{j(s)}(s), \mathcal{B}_1(s), \mathcal{M}(s)\} \\ &= \{T_{S_2} =: s_{S_2,1}, s_{S_2,2}, \dots, s_{S_2,|S_2(T)|}\}, \end{aligned}$$

where  $s_{S_m,k}$  implies the round when the event occurs  $k$ -th time for  $m = 1, 2$ , respectively.

Similarly, let us define a subset of rounds with any fixed  $T > T_M$

$$\begin{aligned} S_0(T) &:= \left\{ s \in [T_M, T] \cap \mathbb{N} : \{\mathcal{B}_1(s), \mathcal{M}^c(s)\} \cup \{\mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s), \mathcal{M}(s)\} \right. \\ &\quad \cup \{m(s) = 1, i(s) = 1, \mathcal{B}_1(s), \mathcal{A}_{j(s)}^c(s), \mathcal{M}(s)\} \\ &\quad \left. \cup \{m(s) \neq 1, i(s) = j(s), \mathcal{B}_1(s), \mathcal{A}_{m(s)}^c(s), \mathcal{B}_{j(s)}(s), \mathcal{M}(s)\} \right\} \end{aligned}$$

and a random variable

$$\begin{aligned} T_S &:= T_M + \sum_{s=T_M+1}^T \mathbb{1}[\mathcal{B}_1(s), \mathcal{M}^c(s)] + \mathbb{1}[\mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s), \mathcal{M}(s)] \\ &\quad + \mathbb{1}[m(s) = 1, i(s) = 1, \mathcal{B}_1(s), \mathcal{A}_{j(s)}^c(s), \mathcal{M}(s)] \\ &\quad + \mathbb{1}[m(s) \neq 1, i(s) = j(s), \mathcal{B}_1(s), \mathcal{B}_{m(s)}^c(s), \mathcal{B}_{j(s)}(s), \mathcal{M}(s)], \end{aligned}$$

such that  $T_S = |S_0(T)| + T_M$  holds.

**First objective** Here, we first aim to show that for  $t \geq T_M$ , it holds

$$1 = \mathbb{1}[t \in S_0(T)] + \mathbb{1}[t \in S_1(T)] + \mathbb{1}[t \in S_2(T)].$$

Since  $\mathcal{B}_1(s)$  always holds for  $s \geq T_M$ , it holds that

$$\begin{aligned}
1 &= \mathbb{1}[\mathcal{B}_1(s)] \\
&= \mathbb{1}[\mathcal{M}^c(s), \mathcal{B}_1(s)] + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s)] \\
&= \mathbb{1}[\mathcal{M}^c(s), \mathcal{B}_1(s)] + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1] + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1] \\
&= \mathbb{1}[\mathcal{M}^c(s), \mathcal{B}_1(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = 1] + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = j(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = m(s), \mathcal{B}_1(s), \mathcal{B}_{m(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = j(s), \mathcal{B}_1(s), \mathcal{A}_{m(s)}^c(s)] \tag{5.22}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{1}[\mathcal{M}^c(s), \mathcal{B}_1(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = 1, \mathcal{A}_{j(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = 1, \mathcal{A}_{j(s)}(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = j(s), \mathcal{B}_{j(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = j(s), \mathcal{B}_{j(s)}(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = m(s), \mathcal{B}_{m(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = j(s), \mathcal{A}_{m(s)}^c(s), \mathcal{B}_{j(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = j(s), \mathcal{A}_{m(s)}^c(s), \mathcal{B}_{j(s)}(s)] \tag{5.23}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{1}[\mathcal{M}^c(s), \mathcal{B}_1(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = 1, \mathcal{A}_{j(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = 1, \mathcal{A}_{j(s)}(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = j(s), \mathcal{B}_{j(s)}(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = j(s), \mathcal{A}_{m(s)}^c(s), \mathcal{B}_{j(s)}(s)] \\
&= \mathbb{1}[s \in S_0(T)] + \mathbb{1}[s \in S_1(T)] + \mathbb{1}[s \in S_2(T)],
\end{aligned}$$

where (5.22) and (5.23) hold from

$$\mathbb{1}[m(s) \neq 1, \mathcal{B}_1(s)] = \mathbb{1}[m(s) \neq 1, \mathcal{B}_1(s), \mathcal{B}_{m(s)}^c(s)] = \mathbb{1}[m(s) \neq 1, \mathcal{B}_1(s), \mathcal{A}_{m(s)}^c(s)]. \tag{5.24}$$

The last equality holds from

$$\begin{aligned}
&\mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s)] \\
&= \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s), m(s) = 1] + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s), m(s) \neq 1] \\
&= \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s), m(s) = 1, i(s) = j(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s), m(s) \neq 1, i(s) = m(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), \mathcal{B}_{i(s)}^c(s), m(s) \neq 1, i(s) = j(s)] \\
&= \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) = 1, i(s) = j(s), \mathcal{B}_{j(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = m(s), \mathcal{B}_{m(s)}^c(s)] \\
&\quad + \mathbb{1}[\mathcal{M}(s), \mathcal{B}_1(s), m(s) \neq 1, i(s) = j(s), \mathcal{A}_{m(s)}^c(s), \mathcal{B}_{j(s)}^c(s)], \tag{5.25}
\end{aligned}$$

where we used (5.24) in (5.25) again. This implies that if  $T \geq T_M$ , then  $[T_M, T] \cap \mathbb{N} = S_0(T) \cup S_1(T) \cup S_2(T)$  holds. Note that if  $s = T_M \geq KD$ , there exists at least one arm  $a \in [K]$  satisfying  $N_a(s) \geq D$ .

**(1) If  $N_1(s) \geq D$ :**

Recall the definition  $T_{S_1} = \inf S_1(T)$  and  $T_{S_2} = \inf S_2(T)$ , which implies the first round when the events in Lemmas 5.6 and 5.7 occur, respectively.

**(1-i)  $S_0(T)$  is a subinterval:** If  $S_0(T)$  consists of consecutive natural numbers, i.e., the subinterval in  $[T_M, T] \cap \mathbb{N}$ , then  $\min(T_{S_1}, T_{S_2}) \leq T_S + 1$  holds since we can only observe events in  $S_1(T)$  or  $S_2(T)$  for  $s > T_S$ .

**(1-ii)  $S_0(T)$  is not a subinterval:** If  $S_0(T)$  is not a subinterval of  $[T_M, T] \cap \mathbb{N}$ , this directly implies that  $\min(T_{S_1}, T_{S_2}) \leq T_S$  from  $[T_M, T] \cap \mathbb{N} = S_0(T) \cup S_1(T) \cup S_2(T)$ .

**(1-iii) Summary:** What we have shown is  $\min(T_{S_1}, T_{S_2}) \leq T_S + 1$ . Let us consider the case  $T_{S_1} < T_{S_2}$ . From the definition of  $T_{S_1}$  where  $i(t) = j(t)$ , we have for  $j(t) = j$  and  $a \neq 1, j$  that

$$\begin{aligned} (N_1(T_{S_1}) + N_j(T_{S_1}))d(\hat{\mu}_1(T_{S_1}), \hat{\mu}_{1,j}(T_{S_1})) \\ \leq N_1(T_{S_1})d(\hat{\mu}_1(T_{S_1}), \hat{\mu}_{1,j}(T_{S_1})) + N_j(T_{S_1})d(\hat{\mu}_j(T_{S_1}), \hat{\mu}_1(T_{S_1})) \\ = T_{S_1}f_j(\mathbf{w}^t; \hat{\mu}(T_{S_1})) \\ \leq T_{S_1}f_a(\mathbf{w}^t; \hat{\mu}(T_{S_1})) \\ \leq N_a(T_{S_1})d(\hat{\mu}_a(T_{S_1}), \hat{\mu}_1(T_{S_1})). \end{aligned}$$

From the assumption  $N_1(T_{S_1}) \geq D$ , it holds that

$$\begin{aligned} N_1(T_{S_1})d(\hat{\mu}_1(T_{S_1}), \hat{\mu}_{1,j}(T_{S_1})) &\geq N_1(T_{S_1})d(\hat{\mu}_1(T_{S_1}), \hat{\mu}_{1,j}(T_{S_1})) \frac{D_1}{\min_{i \in [K]d_i}} \\ &\geq D_1 = \max_{i \neq 1} D_i. \end{aligned}$$

Therefore,

$$\max_{i \in [K]} D_i < \min_{i \neq 1} N_a(T_{S_1})d(\hat{\mu}_a(T_{S_1}), \hat{\mu}_1(T_{S_1})). \quad (5.26)$$

Recall the definition  $D_i = \sup_t \mathbb{1}[\mathcal{B}_i^c(t)]N_i(t)d(\hat{\mu}_i(t), \hat{\mu}_1(t))$ . Thus (5.26) implies that  $\mathcal{B}_a(t)$  holds for all  $t \geq T_{S_1}$  and any  $i \in [K]$ , i.e.,  $T_B \leq T_{S_1} \leq T_S + 1$ . When  $T_{S_2} < T_{S_1}$  holds,  $T_B \leq T_{S_2} \leq T_S + 1$  can be directly derived from Lemma 5.7.

**(2) If  $N_i(s) \geq D$  for  $i \neq 1$ :**

From (1), one can expect that  $T_B$  will be bounded at least if either  $N_{j(s)}(s)$  or  $N_1(s)$  satisfies the condition in (5.26) for any  $s \leq T$ .

**(2-i)  $j(s) = i$  holds for some  $s \in S_1(T)$ :** In this case, we have for  $a \neq 1, i$

$$N_1(s)d(\hat{\mu}_1(s), \hat{\mu}_{1,i}(s)) + N_i(s)d(\hat{\mu}_i(s), \hat{\mu}_{1,i}(s)) = sf_i < sf_a \leq N_a(s)d(\hat{\mu}_a(s), \hat{\mu}_1(s)),$$

where we denote  $\hat{\mu}_{1,i}^{\mathbf{w}^s}(s)$  by  $\hat{\mu}_{1,i}(s)$  for notational simplicity. From  $N_i(s) \geq D$ ,

$$\max_{a \in [K]} D_a \leq N_i(s)d(\hat{\mu}_i(s), \hat{\mu}_{1,i}(s)) \leq \min_{a \neq 1} N_a(s)d(\hat{\mu}_a(s), \hat{\mu}_1(s)), \quad (5.27)$$

which implies  $T_B \leq s$ .

**(2-ii)  $j(s) \neq a$  holds for all  $s \in S_1(T)$ :** Take arbitrary  $t' \in (T_M, \infty) \cap \mathbb{N}$  and assume that there exists an arm  $j' \neq 1$  and a round  $s' \geq t'$  such that  $\mathbb{1}[\mathcal{B}_{j'}^c(s')] = 1$  holds. Note that whenever  $N_{j(s)}(s) \geq D$  holds, substituting  $a = j(s)$  in (5.27) leads to the same inequality, which implies  $T_B \leq s$ .

**(2-iii) Summary:** Therefore, for every arm  $j \neq 1$ ,  $\sum_{s \in S_1(T)} \mathbb{1}[j(s) = j] \leq D$  should hold since  $\sum_{s \in S_1(T)} \mathbb{1}[j(s) = j] > D$  admits the existence of  $s \in S_1(T)$  such that satisfies (5.27), which contradicts to the assumption of the existence of such  $s'$ . In other words,  $\sum_{s \in S_1(T)} \mathbb{1}[j(s) = j] \leq D$  is a necessary condition to satisfy the assumption of the existence of  $j'$  and  $s'$  satisfying  $\mathbb{1}[\mathcal{B}_{j'}^c(s')] = 1$ . From the definition of  $S_1(T)$ , for any  $s \in S_1(T)$ ,  $N_{j(s)}(s+1) = N_{j(s)}(s) + 1$  holds. Hence, at worst, if  $|S_1(T) \cap [T_M, t']| \geq (K-2)D$  holds at some round  $t'$ , there exists  $s \in S_1(T) \cap [T_M, t']$  such that  $N_{j(s)}(s) \geq D$ . Therefore,  $T_B$  is at most the round until  $S_1(T)$  occur  $(K-2)D$  times.

Similarly, if the event in  $S_2(T)$  occurs  $D$  times at some round  $t''$ , then  $N_1(t'') \geq D$  holds from the sampling rule. This implies that  $B_i(s)$  holds for all  $i \in [K]$  for  $s \geq t''$  from (5.26), i.e.,  $T_B$  is at most the round until  $S_2(T)$  occur  $D$  times.

### (3) Conclusion

In summary, we have  $[T_M, T] \cap \mathbb{N} = S_0(T) \cup S_1(T) \cup S_2(T)$  and there exists an arm  $i$  satisfying  $N_i(t) \geq D$ . If  $N_1(s) \geq D$ , then  $T_B \leq T_S + 1$  holds. If  $N_i(s) \geq D$  holds for  $i \neq 1$ , then  $T_B$  is at most the round  $s_{S_1, (K-2)D}$  when the event in  $S_1(T)$  occurs  $(K-2)D$  times or  $s_{S_2, D}$  when the event in  $S_2(T)$  occur  $D$  times. Hence, we have

$$T_B \leq T_S + (K-2)D + D + 1,$$

where  $T_S = T_M + |S_0(T)| = \max(T_C, KD) + |S_0(T)|$ . Then, we have

$$\begin{aligned} \mathbb{E}[T_B] &\leq \mathbb{E}[T_S] + (K-1)\mathbb{E}[D] + 1 \\ &\leq \mathbb{E}[T_C] + (2K-1)\mathbb{E}\left[\sup_{i \neq 1} \sup_{s \geq t} \mathbb{1}[\mathcal{B}_i^c(s)] N_i(s) d(\hat{\mu}_i(s), \hat{\mu}_1(s))\right] \\ &\quad + \mathbb{E}\left[\sum_{t=T_M}^T \mathbb{1}[\mathcal{M}^c(t)] + \mathbb{1}[m(t) = 1, i(t) = 1, \mathcal{B}_1(t), \mathcal{A}_{j(t)}^c(t), \mathcal{M}(t)] \right. \\ &\quad \left. + \mathbb{1}[m(t) \neq 1, i(t) = j(t), \mathcal{B}_1(t), \mathcal{A}_{m(t)}^c(t), \mathcal{B}_{j(t)}(t), \mathcal{M}(t)] \right. \\ &\quad \left. + \mathbb{1}[\mathcal{B}_1(t), \mathcal{B}_{i(t)}^c(t), \mathcal{M}(t)]\right] + 1. \end{aligned}$$

Then, the following five lemmas conclude the proofs. □

**Lemma 5.8.** *For a bounded region of parameters  $R \subset \mathbb{R}$ , it holds that for arbitrary  $\mu' \in R$  and  $i \in [K]$*

$$\mathbb{E}\left[\sup_{n \in \mathbb{N}, \mu' \in R} \mathbb{1}[|\hat{\mu}_{i,n} - \mu_i| \geq \epsilon] n d(\hat{\mu}_{i,n}, \mu')\right] = \mathcal{O}(d_\epsilon^{-1}),$$

where  $\hat{\mu}_{i,n}$  is the empirical mean reward of the arm  $i$  when it is played  $n$  times.

Here, note that  $\hat{\mu}_{i,n}$  is different from  $\hat{\mu}_{a,b}(t)$  that denotes the weighted average of their empirical mean. Lemma 5.8 provides the finiteness of the expectation of  $D_i$  for any  $i \in [K]$ .

**Lemma 5.9.** *For the finite number of arms  $K$  and any  $T \in \mathbb{N}$ , it holds that*

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=1}^T \mathbb{1} \left[ m(t) = 1, i(t) = 1, \mathcal{B}_1(t), \mathcal{A}_{j(t)}^c(t), \mathcal{M}(t) \right] \right] &\leq \mathcal{O}(K d_\epsilon^{-1}), \\ \mathbb{E} \left[ \sum_{t=1}^T \mathbb{1} \left[ i(t) = j(t), \mathcal{A}_{m(t)}^c(t), \mathcal{B}_{j(t)}(t), \mathcal{M}(t) \right] \right] &\leq \mathcal{O}(K^2 d_\epsilon^{-1}). \end{aligned}$$

**Lemma 5.10.** *For the finite number of arms  $K$  and any  $T \in \mathbb{N}$ , it holds that*

$$\mathbb{E} \left[ \sum_{t=1}^T \mathbb{1} \left[ \mathcal{B}_{i(t)}^c(t), \mathcal{M}(t) \right] \right] \leq \mathcal{O}(K d_\epsilon^{-1}).$$

The proofs of Lemmas 5.8–5.10 are provided in Section 5.6.4.

**Lemma 5.11.** *For the finite number of arms  $K$  and any  $T \in \mathbb{N}$ , it holds that*

$$\mathbb{E} \left[ \sum_{t=1}^T \mathbb{1}[\mathcal{M}^c(t)] \right] \leq \mathcal{O}(K^2 d_\epsilon^{-2}).$$

The proof of Lemma 5.11 is given in Section 5.6.5.

**Lemma 5.12.** *Under Algorithm 5, it holds for any  $\epsilon \in (0, \frac{\mu_1 - \mu_2}{2})$  that*

$$\mathbb{E}[T_C] \leq C(\pi_j, \boldsymbol{\mu}, \epsilon) + 4d_\epsilon^{-3},$$

where  $C(\pi_j, \boldsymbol{\mu}, \epsilon)$  specified in Lemma 5.15.

The proof of Lemma 5.12 is given in Section 5.6.6, where we adapt the analysis in Korda et al. [2013] to our problem.

### 5.6.3 Proofs of technical lemmas for Theorem 5.3: Sufficient conditions for the convergence of estimates

Here, we provide the proof of Lemmas 5.6 and 5.7.

*Proof of Lemma 5.6.* Since  $i(t) = j$  implies

$$d(\hat{\mu}_j(t), \hat{\mu}_{1,j}(t)) \geq d(\hat{\mu}_1(t), \hat{\mu}_{1,j}(t)),$$

we have

$$d(\hat{\mu}_j(t), \hat{\mu}_{1,j}(t)) \geq \underline{d}_j,$$

from the definition of  $\underline{d}_j$  in (5.17).

Then, we have

$$\begin{aligned} t f_j(\mathbf{w}^t, \hat{\boldsymbol{\mu}}(t)) &= N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_{1,j}(t)) + N_j(t) d(\hat{\mu}_j(t), \hat{\mu}_{1,j}(t)) \\ &\geq N_j(t) \underline{d}_j > D_1 \end{aligned}$$

On the other hand, if  $|\hat{\mu}_1(t) - \mu_1| \geq \epsilon$  and  $|\hat{\mu}_j(t) - \mu_j| \leq \epsilon$ , then

$$t f_j(\mathbf{w}^t, \hat{\boldsymbol{\mu}}(t)) \leq N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_j(t)) \leq D_1$$

by the definition of  $D_1 = \sup_{i \neq 1} D_i$ . Therefore,  $|\hat{\mu}_1(t) - \mu_1| \geq \epsilon$  cannot hold.

Under  $|\hat{\mu}_1(t) - \mu_1| \leq \epsilon$  and  $|\hat{\mu}_j(t) - \mu_j| \leq \epsilon$ , we see that

$$\begin{aligned} N_j(t) \underline{d}_j &\leq t f_j(\mathbf{w}^t, \hat{\boldsymbol{\mu}}(t)) \leq N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_j(t)) \\ &\leq N_1(t) d(\mu_1 + \epsilon, \mu_j - \epsilon), \end{aligned}$$

which completes the proof.  $\square$

*Proof of Lemma 5.7.* Since  $j(t) = \arg \min_{i \neq m(t)} t f_i(\mathbf{w}^t, \hat{\boldsymbol{\mu}}(t))$  and  $i(t) = 1$ , it holds for all  $i \neq 1$  that

$$t f_i(\mathbf{w}^t, \hat{\boldsymbol{\mu}}(t)) \geq t f_{j(t)}(\mathbf{w}^t, \hat{\boldsymbol{\mu}}(t))$$

and

$$d(\hat{\mu}_1(t), \hat{\mu}_{1,j(t)}(t)) \geq d(\hat{\mu}_{j(t)}(t), \hat{\mu}_{1,j(t)}(t)).$$

Then, we can use the same argument as Lemma 5.6 by exchanging the role of 1 and  $j$ .  $\square$

#### 5.6.4 Proofs of technical lemmas for Theorem 5.3: Boundedness of the number of rounds where estimates do not converge

Here, we provide the proof of Lemmas 5.8–5.10. Firstly, to prove Lemma 5.8, we require the lemma below, whose proof is postponed to Section 5.6.7.

**Lemma 5.13.** *Let  $R \subset \mathbb{R}$  be a bounded region of parameters and fix arbitrary  $\mu_0$ . Then, there exists  $a, b \geq 0$  such that*

$$d(\mu, \mu') \leq ad(\mu, \mu_0) + b$$

for arbitrary  $\mu \in \mathbb{R}$  and  $\mu' \in R$ .

*Proof of Lemma 5.8.* Let  $P(z) := \mathbb{P}[d(\hat{\mu}_{i,n}, \mu_i) \geq z]$ . Then, by Chernoff bound, we have  $P(z) \leq 2e^{-nz}$ . Therefore,

$$\begin{aligned} \mathbb{E} \left[ \mathbb{1}[|\hat{\mu}_{i,n} - \mu_i| \geq \epsilon] \sup_{\mu' \in R} d(\hat{\mu}_{i,n}, \mu') \right] &\leq \mathbb{E}[\mathbb{1}[|\hat{\mu}_{i,n} - \mu_i| \geq \epsilon](ad(\hat{\mu}_{i,n}, \mu_i) + b)] \\ &\leq 2be^{-nd_\epsilon} + a \int_{d_\epsilon}^{\infty} z d(-P(z)) \\ &= 2be^{-nd_\epsilon} + a \left( -[zP(z)]_{d_\epsilon}^{\infty} + \int_{d_\epsilon}^{\infty} zP(z) dz \right) \\ &\leq 2be^{-nd_\epsilon} + 2ad_\epsilon e^{-nd_\epsilon} + a \int_{d_\epsilon}^{\infty} zP(z) dz \\ &\leq 2be^{-nd_\epsilon} + 2ad_\epsilon e^{-nd_\epsilon} + 2a \left[ -\frac{ze^{-nz}}{n} - \frac{e^{-nz}}{n^2} \right]_{d_\epsilon}^{\infty} \\ &\leq 2 \left( b + a \left( d_\epsilon + \frac{d_\epsilon}{n} + \frac{1}{n^2} \right) \right) e^{-nd_\epsilon}, \end{aligned}$$

where  $d_\epsilon := \min_{i \in [K]} \{d(\mu_i - \epsilon, \mu_i), d(\mu_i + \epsilon, \mu_i)\}$  and the first inequality holds from Lemma 5.13. Since this quality decays exponentially in  $n$ , it is straightforward that

$$\begin{aligned} \mathbb{E} \left[ \sup_{n \in \mathbb{N}, \mu' \in R} \mathbb{1}[|\hat{\mu}_{i,n} - \mu_i| \geq \epsilon] nd(\hat{\mu}_{i,n}, \mu') \right] \\ \leq \sum_{n=1}^{\infty} \mathbb{E} \left[ \mathbb{1}[|\hat{\mu}_{i,n} - \mu_i| \geq \epsilon] \sup_{\mu' \in A} d(\hat{\mu}_{i,n}, \mu') \right] = \mathcal{O}(d_\epsilon^{-1}). \end{aligned}$$

$\square$



*Proof of Lemma 5.9.* For  $j(t) = j$ , we first consider

$$D_j = \sup_t \{ \mathbb{1} [ | \hat{\mu}_j(t) - \mu_i | \geq \epsilon ] N_j(t) d(\hat{\mu}_j(t), \hat{\mu}_1(t)) \}.$$

Note that on  $\mathcal{B}_1(t)$ ,  $\hat{\mu}_1(t) \in [\mu_1 - \epsilon, \mu_1 + \epsilon]$  is bounded so that we can apply Lemmas 5.8 and 5.13. We first show the existence of a bounded constant  $c_j^* \in \mathbb{R}_+$  such that

$$N_1(t) \leq c_j^* D_j,$$

where

$$c_j^* = \min \left( c_j, \frac{x'_j}{d_\zeta} \right)$$

for constants  $c_j$ ,  $x'_j$  and  $d_\zeta$  that depend on models.

**(1) When  $\hat{\mu}_j(t) \not\approx \hat{\mu}_{m(t)}(t)$**

From their definitions, we have

$$0 \leq N_j(t) d(\hat{\mu}_i(t), \hat{\mu}_{1,j}(t)) \leq N_j(t) d(\hat{\mu}_j(t), \hat{\mu}_1(t)) \leq D_i$$

and

$$\begin{aligned} N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_{1,j}(t)) &\leq N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_{1,j}(t)) + N_j(t) d(\hat{\mu}_i(t), \hat{\mu}_{1,j}(t)) \\ &= tg(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)). \end{aligned}$$

Let us consider

$$\psi(x; t) = xd(\hat{\mu}_{m(t)}(t), \hat{\mu}_{m(t),j}(x; t)) + d(\hat{\mu}_j(t), \hat{\mu}_{m(t),j}(x; t)),$$

where  $\hat{\mu}_{a,b}(x; t) = \frac{x\hat{\mu}_a(t) + \hat{\mu}_b(t)}{x+1}$ . One can see that  $\psi(x; t)$  is strictly increasing with respect to  $x$  since  $\psi'(x; t) = d(\hat{\mu}_{m(t)}(t), \hat{\mu}_{m(t),j}(x; t)) > 0$  and it tends to  $d(\hat{\mu}_j(t), \hat{\mu}_{m(t)}(t))$  when  $x$  goes to infinity [Garivier and Kaufmann, 2016]. Then, under the condition  $\{m(t) = 1, j(t) = j\}$ , it holds that

$$\begin{aligned} tg(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)) &= N_j(t) \psi \left( \frac{N_1(t)}{N_j(t)}; t \right) \leq N_j(t) d(\hat{\mu}_j(t), \hat{\mu}_1(t)) \\ &\leq D_j. \end{aligned}$$

Therefore,

$$N_1(t) \leq \frac{1}{d(\hat{\mu}_1(t), \hat{\mu}_{1,j}(t))} D_j.$$

Note that there exists a constant  $c_j$  such that  $\frac{1}{d(\hat{\mu}_1(t), \hat{\mu}_{1,j}(t))} \leq c_j < \infty$  when  $\hat{\mu}_a(t) \not\approx \hat{\mu}_{m(t)}(t)$ , which shows the existence of  $c_j^*$ .

**(2) When  $\hat{\mu}_j(t) \approx \hat{\mu}_{m(t)}(t)$**

Here,  $i(t) = 1$  implies that

$$d(\hat{\mu}_1(t), \hat{\mu}_{1,j}^{\mathbf{w}^t}(t)) \geq d(\hat{\mu}_j(t), \hat{\mu}_{1,j}^{\mathbf{w}^t}(t)). \quad (5.28)$$

Note that as  $\frac{w_1(t)}{w_j(t)}$  increases, RHS of (5.28) decreases and LHS of (5.28) increases simultaneously. Therefore,

$$\forall t \in \mathbb{N}, \exists x_{j,t}^* \in \mathbb{R}_+ \text{ s.t. } \frac{w_1(t)}{w_j(t)} = x_{j,t}^* \Leftrightarrow d(\hat{\mu}_1(t), \hat{\mu}_{1,j}^{\mathbf{w}^t}(t)) = d(\hat{\mu}_j(t), \hat{\mu}_{1,j}^{\mathbf{w}^t}(t)).$$

Note that  $x_{j,t}^*$  depends on the distribution of reward and history  $H_t$  until round  $t$ , e.g.,  $\forall t \in \mathbb{N}$ ,  $x_{j,t}^* = 1$  for the Gaussian distribution. Since  $\hat{\mu}_1(t)$  is bounded under  $\{\mathcal{B}_1(t)\}$  and  $\hat{\mu}_j(t) \in (\mu_j + \epsilon, \hat{\mu}_1(t)] \subset (\mu_j + \epsilon, \mu_1 + \epsilon]$  holds under  $\{\mathcal{B}_1(t), \mathcal{A}_j^c(t), m(t) = 1\}$ , there exists  $x'_j \in \mathbb{R}_+$  such that for any  $t \in \mathbb{N}$

$$N_1(t) > x'_j N_j(t) \implies d(\hat{\mu}_1(t), \hat{\mu}_{1,j}(t)) < d(\hat{\mu}_j(t), \hat{\mu}_{1,j}(t)), \text{ i.e., } i(t) = j.$$

Let consider a bounded region  $R = [\mu_1 - \epsilon, \mu_1 + \epsilon] \subset \mathbb{R}$  and a random variable

$$D_j = \sup_{t \in \mathbb{N}} \sup_{\mu' \in A} \left\{ \mathbb{1}[|\hat{\mu}_j(t) - \mu_j| \geq \epsilon] N_j(t) d(\hat{\mu}_j(t), \mu') \right\}, \quad j \in [K] \setminus \{1\}.$$

Since  $m(t) = 1$  holds under the condition, we have

$$\sup_{\mu' \in A} d(\hat{\mu}_j(t), \mu') = \max\{d(\hat{\mu}_j(t), \mu_1 - \epsilon), d(\hat{\mu}_j(t), \mu_1 + \epsilon)\}$$

and  $\hat{\mu}_1(t) > \hat{\mu}_j(t)$ . Let  $\zeta(\epsilon) \in A$  be a point such that  $d(\zeta, \mu_1 - \epsilon) = d(\zeta, \mu_1 + \epsilon) = d_\zeta$ . Then, it holds that

$$\sup_{\mu' \in A} d(\hat{\mu}_j(t), \mu') > d_\zeta.$$

Note that  $d_\zeta$  and  $x'_j$  only depend on the models. Therefore, there exists a constant  $c_j^* \in \mathbb{R}_+$  such that

$$N_1(t) \leq \frac{x'_j}{d_\zeta} D_j \leq c_j^* D_j.$$

### (3) Conclusion

From Lemma 5.8, we obtain

$$\begin{aligned} \mathbb{E} \left[ \sum_{i \in [K] \setminus \{1\}} \sum_{t=1}^{\tau} \mathbb{1} \left[ m(t) = 1, i(t) = 1, \mathcal{B}_1(t), j(t) = i, \mathcal{A}_{j(t)}^c(t), \mathcal{M}(t) \right] \right] \\ \leq \mathbb{E} \left[ \sum_{i \in [K] \setminus \{1\}} \sum_{t=1}^{\infty} \mathbb{1} [i(t) = 1, N_1(t) \leq c_i^* D_i] \right] \\ \leq \sum_{i \in [K] \setminus \{1\}} c_i^* \mathbb{E}[D_i] \leq \mathcal{O}(K d_\epsilon^{-1}), \end{aligned}$$

which concludes the first case.

Similarly, the second case can be bounded by considering  $R_j = [\mu_j - \epsilon, \mu_j + \epsilon]$  and

$$D_{m(t),j} = \sup_n \sup_{\mu' \in R_j} \left\{ \mathbb{1}[|\hat{\mu}_{m(t)}(n) - \mu_{m(t)}| \geq \epsilon] n d(\hat{\mu}_{m(t)}(n), \mu') \right\}$$

for every  $m(t) \in [K]$  and  $j \in [K] \setminus \{m(t)\}$ . Since  $\hat{\mu}_j(t) \in R_j$  holds under  $\{\mathcal{B}_j(t)\}$ , we can apply Lemmas 5.8 and 5.13 by exchanging the role of  $m(t)$  and  $j$ , which concludes the proof.  $\square$

*Proof of Lemma 5.10.* From the Chernoff bound, it holds for any arm  $i \in [K]$  that

$$\mathbb{P}[|\hat{\mu}_i(t) - \mu_i| \geq \epsilon | N_i(t) = n] \leq 2e^{-nd_\epsilon}, \quad (5.29)$$

where  $d_\epsilon$  is defined in (5.9). One can rewrite the expectation as

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=1}^T \mathbb{1} \left[ \mathcal{B}_{i(t)}^c(t), \mathcal{M}(t) \right] \right] &= \mathbb{E} \left[ \sum_{i=1}^K \sum_{t=1}^T \sum_{n=1}^{\infty} \mathbb{1} \left[ i(t) = i, \mathcal{B}_{i(t)}^c(t), \mathcal{M}(t), N_{i(t)}(t) = n \right] \right] \\ &= \mathbb{E} \left[ \sum_{i=1}^K \sum_{t=1}^T \sum_{n=1}^{\infty} \mathbb{1} \left[ i(t) = i, \mathcal{B}_i^c(t), \mathcal{M}(t), N_i(t) = n \right] \right] \end{aligned}$$

For every arm  $i \in [K]$ , an event  $\{i(t) = i, N_i(t) = n\}$  could happen at most once for any  $n \in \mathbb{N}$ . Therefore, by applying (5.29), one has

$$\mathbb{E} \left[ \sum_{t=1}^T \mathbb{1} \left[ \mathcal{B}_{i(t)}^c(t), \mathcal{M}(t) \right] \right] \leq \sum_{i=1}^K \sum_{n=1}^{\infty} 2e^{-nd_\epsilon} \leq \mathcal{O}(Kd_\epsilon^{-1}),$$

which concludes the proof.  $\square$

### 5.6.5 Proof of technical lemma for Theorem 5.3: An upper bound on the number of rounds where TE occurs

Here, we provide the proof of Lemma 5.11, which shows that the expected number of rounds where Thompson samples and the empirical mean estimates disagree is finite. Before beginning the proof, we present the posterior concentration result when we employ the Jeffreys prior in the SPEF.

**Lemma 5.14** (Theorem 4 in Korda et al. [2013]). *For the Jeffreys prior and  $d_\epsilon$  defined in (5.9), there exists constants  $C_{1,a} = C_1(\theta_a, A) > 0$ ,  $C_{2,a} = C_2(\theta_a, A, \epsilon) > 0$  and  $N(\theta_a, A)$  such that for any  $N_a(t) \geq N(\theta_a, A)$ ,*

$$\mathbb{1}[\mathcal{B}_a(t)] \mathbb{P}[\tilde{\mathcal{B}}_a^c(t) | X_{a,N_a(t)}] \leq 2C_{1,a} N_a(t) e^{-(N_a(t)-1)(1-\epsilon C_{2,a})d_\epsilon}$$

whenever  $\epsilon$  is such that  $1 - \epsilon C_{2,a}(\epsilon) > 0$ . Note that  $A$  is a convex function in (5.1).

*Proof of Lemma 5.11.* Let us define  $L(\theta) := \frac{1}{2} \min(\sup_y p(y|\theta), 1)$  and an event

$$\tilde{E}_a(t) = \left( \exists 1 \leq s' \leq N_a(t) : p(x_{a,s'} | \theta_a) \geq L(\theta_a), \left| \frac{\sum_{s=1, s \neq s'}^{N_a(t)} x_{a,s}}{N_a(t) - 1} - \mu_a \right| \leq \epsilon \right).$$

Consider

$$\begin{aligned} \sum_{t=1}^T \mathbb{1}[\mathcal{M}^c(t)] &= \sum_{t=1}^T \sum_{i \in [K]} \mathbb{1}[i(t) = i, \mathcal{M}^c(t)] \\ &= \sum_{t=1}^T \sum_{i \in [K]} \mathbb{1}[i(t) = i, \tilde{E}_a^c(t), \mathcal{M}^c(t)] + \mathbb{1}[i(t) = i, \tilde{E}_a(t), \mathcal{M}^c(t)] \end{aligned}$$

It is shown by Korda et al. [2013] that

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=1}^T \mathbb{1}[i(t) = i, \tilde{E}_i^c(t), \mathcal{M}^c(t)] \right] &\leq \sum_{t=1}^{\infty} \mathbb{P}(p(x_{i,1} | \theta_a) \leq L(\theta_a))^t + \sum_{t=1}^{\infty} 2te^{-(t-1)d_\epsilon} \\ &\leq \mathcal{O}(d_\epsilon^{-2}). \end{aligned} \tag{5.30}$$

Then, consider

$$\begin{aligned} \sum_{t=1}^T \mathbb{1}[i(t) = i, \tilde{E}_i(t), \mathcal{M}^c(t)] &= \sum_{t=1}^T \left( \mathbb{1}[i(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \right. \\ &\quad \left. + \mathbb{1}[i(t) = i, \tilde{\mathcal{B}}_i^c(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \right). \end{aligned}$$

On  $\tilde{E}_i(t)$ , the following holds for a constant  $N(\theta_i, A)$  from Lemma 5.14.

$$\begin{aligned}
& \mathbb{E} \left[ \sum_{t=1}^T \sum_{i \in [K]} \mathbb{1}[i(t) = i, \tilde{\mathcal{B}}_i^c(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \right] \\
& \leq \sum_{i \in [K]} N(\theta_i, A) + \sum_{i \in [K]} \sum_{\substack{t: i(t)=i \\ N_a(t) \geq N(\theta_i, A)}}^T 2C_{1,i} e^{-(N_i(t)-1)(1-\epsilon C_{2,i})d_\epsilon + \log(N_i(t))} \\
& \leq \sum_{i \in [K]} N(\theta_i, A) + \sum_{i \in [K]} \sum_{n=N(\theta_i, A)}^{\infty} 2C_{1,i} n e^{-(n-1)(1-\epsilon C_{2,i})d_\epsilon} \\
& \leq \mathcal{O}(Kd_\epsilon^{-2}),
\end{aligned}$$

where the second inequality holds since  $N_i(t)$  increases when  $\{i(t) = i\}$  happens.

Finally, we will show that

$$\sum_{t=1}^T \sum_{i \in [K]} \mathbb{1}[i(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \leq \mathcal{O}(K^2 d_\epsilon^{-2}).$$

On  $\mathcal{M}^c(t)$ ,  $i(t) \in \{m(t), \tilde{m}(t)\}$  holds so that

$$\begin{aligned}
& \sum_{t=1}^T \sum_{i \in [K]} \mathbb{1}[i(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \\
& \leq \sum_{t=1}^T \sum_{i \in [K]} \mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \\
& \quad + \sum_{t=1}^T \sum_{i \in [K]} \mathbb{1}[i(t) = \tilde{m}(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t)].
\end{aligned}$$

Let us define  $N_A = \max_{a \in [K]} N(\theta_a, A)$ . For any  $i \in [K]$ , we have

$$\begin{aligned}
& \sum_{t=1}^T \mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \\
& \leq N_A + \sum_{t=1}^T \mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t), N_i(t) \geq N_A]
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{t=1}^T \mathbb{1}[i(t) = \tilde{m}(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t)] \\
& \leq N_A + \sum_{t=1}^T \mathbb{1}[i(t) = \tilde{m}(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t), N_i(t) \geq N_A].
\end{aligned}$$

Consider

$$\begin{aligned}
& \mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t), N_i(t) \geq N_A] = \\
& \sum_{j \in [K] \setminus \{i\}} \underbrace{\mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t), N_i(t) \geq N_A, \tilde{m}(t) = j, \tilde{E}_j(t)]}_{(*)} \\
& \quad + \underbrace{\mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t), N_i(t) \geq N_A, \tilde{m}(t) = j, \tilde{E}_j^c(t)]}_{(\star)}.
\end{aligned}$$

Similarly to (5.30), it holds that  $\mathbb{E}[\sum_t(\star)] \leq \mathcal{O}(d_\epsilon^{-2})$ . On  $\mathcal{M}^c(t)$ ,  $\{i(t) = m(t)\}$  implies that  $\{N_{m(t)}(t) \leq N_{\tilde{m}(t)}(t)\}$ , i.e.,  $N_j(t) \geq N_i(t) \geq N_A$  so that one can apply Lemma 5.14. Hence,

$$\sum_t \mathbb{E}[(*)] \leq \mathcal{O}(d_\epsilon^{-2}) + \sum_t \mathbb{E} \left[ \mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t)] \cdot \mathbb{1}[\mathcal{M}^c(t), N_i(t) \geq N_A, \tilde{m}(t) = j, \tilde{E}_j(t), \tilde{\mathcal{B}}_j(t)] \right].$$

From its definition, on  $\tilde{E}_i(t)$ , the empirical mean reward of arm  $i$  is well concentrated around its true mean. Thus,

$$m(t) = i, \tilde{E}_i(t), \tilde{E}_j(t) \implies i > j.$$

However, on  $\{\tilde{\mathcal{B}}_i(t), \tilde{\mathcal{B}}_j(t), \tilde{m}(t) = j\}$ ,  $i < j$  holds, which is a contradiction. Therefore,

$$\mathbb{1}[i(t) = m(t) = i, \tilde{\mathcal{B}}_i(t), \tilde{E}_i(t), \mathcal{M}^c(t), N_i(t) \geq N_A, \tilde{m}(t) = j, \tilde{E}_j(t), \tilde{\mathcal{B}}_j(t)] = 0,$$

which leads to

$$\mathbb{E} \left[ \sum_{t=1}^T \mathbb{1}[\mathcal{M}^c(t)] \right] = \mathcal{O}(K^2 d_\epsilon^{-2}).$$

□

### 5.6.6 Proof of technical lemma for Theorem 5.3: Analysis with TS

Here, we provide the proof of Lemma 5.12.

*Proof.* Let us define an event

$$\mathcal{C}(t) := \bigcup_{s=t}^{\infty} \{\mathcal{B}_1^c(s)\}$$

so that  $\mathcal{C}^c(t) = \bigcap_{s=t}^{\infty} \{\mathcal{B}_1(s)\}$  implies only  $\mathcal{B}_1(s)$  occurs for  $s \geq t$ , meaning that  $\mathcal{C}(t) \Leftrightarrow \{T_C \geq t\}$ . Therefore,

$$\begin{aligned} \mathbb{E}[T_C] &= \sum_{s=1}^{\infty} \mathbb{P}[T_C \geq s] = \sum_{s=1}^{\infty} \mathbb{P}[\mathcal{C}(s)] \\ &= \sum_{s=1}^{\infty} \mathbb{P}[\mathcal{C}(s), N_1(s) \leq \sqrt{s}] + \mathbb{P}[\mathcal{C}(s), N_1(s) \geq \sqrt{s}]. \end{aligned}$$

From the Chernoff bound, we can derive the upper bound of the second term as

$$\begin{aligned} \sum_{s=1}^{\infty} \mathbb{P}[\mathcal{C}(s), N_1(s) \geq \sqrt{s}] &\leq \sum_{s=1}^{\infty} \sum_{n=\sqrt{s}}^{\infty} \mathbb{P}[|\hat{\mu}_{1,n} - \mu_1| \geq \epsilon] \\ &\leq \sum_{s=1}^{\infty} \sum_{n=\sqrt{s}}^{\infty} 2e^{-nd_\epsilon} \\ &\leq \sum_{s=1}^{\infty} \frac{2}{d_\epsilon} e^{-\sqrt{s}d_\epsilon} \\ &\leq \frac{2}{d_\epsilon} \int_0^{\infty} e^{-\sqrt{s}d_\epsilon} ds = \frac{2}{d_\epsilon} \int_0^{\infty} 2xe^{-d_\epsilon x} dx \\ &= 4d_\epsilon^{-3}. \end{aligned}$$

Then, the Lemma 5.15 below concludes the proof. □

**Lemma 5.15.** *For the finite number of arms  $K < \infty$ , and  $\epsilon \in (0, \frac{\mu_1 - \mu_2}{2})$ , there exists some constants  $C(\pi_j, \mu, \epsilon) < \infty$  such that*

$$\sum_{s=1}^{\infty} \mathbb{P}[\mathcal{C}(s), N_1(s) \leq \sqrt{s}] \leq C(\pi_j, \mu, \epsilon).$$

The proof of Lemma 5.15 is given in 5.6.8.

### 5.6.7 Proof of technical lemma for Lemma 5.8

*Proof.* It holds from the expression of KL divergence that

$$\begin{aligned} d(\mu, \mu') - d(\mu, \mu_0) &= A(\theta(\mu_0)) - A(\theta(\mu')) + (\theta(\mu') - \theta(\mu_0))\mu \\ &\leq A(\theta(\mu_0)) - \inf_{x \in R} A(\theta(x)) + |\mu| \sup_{x \in A} |\theta(x) - \theta(\mu_0)|. \end{aligned}$$

Since  $d(\mu, \mu_0)$  is convex with respect to  $\mu$ , there exist constant  $a', b' \geq 0$  such that  $|\mu| \leq a'd(\mu, \mu_0) + b'$ . Letting  $a := 1 + a' \sup_{x \in A} |\theta(x) - \theta(\mu_0)|$  and  $b := b' \sup_{x \in A} |\theta(x) - \theta(\mu_0)| + A(\theta(\mu_0)) - \inf_{x \in A} A(\theta(x))$  concludes the proof.  $\square$

### 5.6.8 Proof of technical lemma for Lemma 5.12

Here, we present the proof of Lemma 5.15, where we adapt the proof techniques considered in Kaufmann et al. [2012b] and Korda et al. [2013]. Before beginning, we introduce some results in Korda et al. [2013].

The following Lemma shows the concentration inequality when an arm is played sufficiently.

**Lemma 5.16** (Lemma 10 in Korda et al. [2013]). *For every  $a \in [K]$  and  $\epsilon > 0$ , there exist constants  $C'_a = C'(\mu_a, \epsilon, A)$  and  $N$  such that for  $t \geq N_K$ ,*

$$\begin{aligned} \mathbb{P}[\exists s \leq t, \exists a \neq 1 : |\hat{\mu}_a(s) - \mu_a| \geq \epsilon, N_a(s) > C'_a \log t] &\leq \frac{2(K-1)}{t^3} \\ \mathbb{P}[\exists s \leq t, \exists a \neq 1 : |\tilde{\mu}_a(s) - \mu_a| \geq \epsilon, N_a(s) > C'_a \log t] &\leq \frac{4(K-1)}{t^3}. \end{aligned}$$

Note that we use the upper bound with the order of  $\mathcal{O}(t^{-3})$  differently from the original lemma whose order is  $\mathcal{O}(t^{-2})$ . This can be done simply by changing the constant term with a multiplication of  $3/2$ .

The following lemma holds for the SPEF.

**Lemma 5.17** (Lemma 9 in Korda et al. [2013]). *There exists a constant  $C = C(\pi_j) < 1$ , such that for every (random) interval  $I$  and for every positive function  $\ell$ , one has*

$$\mathbb{P}[\forall s \in I, \tilde{\mu}_1(s) \leq \mu_2 + \epsilon, |I| \geq \ell(t)] \leq C^{\ell(t)}.$$

*Proof of Lemma 5.15.* Let  $\tau_n$  denote  $n$ -th time when arm 1 is played and  $\xi_n = (\tau_{n+1} - 1) - \tau_n$  be the time between  $n+1$ -th and  $n$ -th time of arm 1 playing. From the definition, it holds that

$$\mathbb{P}[N_1(t) \leq \sqrt{t}, \mathcal{C}(t)] \leq \sum_{n=0}^{\lfloor \sqrt{t} \rfloor} \mathbb{P}[\xi_n \geq \sqrt{t} - 1, \mathcal{C}(t)].$$

For simplicity, let us define an event

$$G_n := \{\xi_n \geq \sqrt{t} - 1, \mathcal{C}(t)\} = \{\xi_n \geq \sqrt{t} - 1, \{\exists n \geq N_1(t) : |\hat{\mu}_{1,n} - \mu_1| \geq \epsilon\}\}$$

so that

$$\mathbb{P}[N_1(t) \leq \sqrt{t}, \mathcal{C}(t)] \leq \sum_{n=0}^{\lfloor \sqrt{t} \rfloor} \mathbb{P}[G_n].$$

On  $G_n$ , we define an index set  $I_n$  and its subset  $I_{n,l}$

$$I_n := [\tau_n, \tau_n + \lceil \sqrt{t} - 1 \rceil] \subset [\tau_n, \tau_{n+1}]$$

$$I_{n,l} := \left[ \tau_n + \left\lfloor \frac{l-1}{K}(\sqrt{t} - 1) \right\rfloor, \tau_n + \left\lfloor \frac{l}{K}(\sqrt{t} - 1) \right\rfloor \right], \quad l \in [K].$$

Note that the inclusion on  $I_n$  holds under  $G_n$ . In the analysis of Thompson sampling [Agrawal and Goyal, 2012, Kaufmann et al., 2012b, Korda et al., 2013], an arm  $a$  is called *saturated* if  $N_a(t) \geq C'_a \log t$  for a constant  $C'_a$  that depends on the model.

In this chapter, we call an arm  $i$  is saturated if  $N_i(t) \geq \max_{a \in [K]} C_a \log t$  for a constant  $C_a$  such that

$$C_a \geq C'_a \frac{d(\mu_2 + \epsilon, \mu_K - \epsilon)}{\underline{d}_a}.$$

Note that  $C_a$ 's are also constants that only depend on the model, and  $C_a \geq C'_a$  holds from the definition of  $\underline{d}_a$ , so that Lemma 5.16 is still applicable. For each interval  $I_n$ , let introduce

- $F_{n,l}$ : the event that by the end of the interval  $I_{n,l}$  at least  $l$  suboptimal arms are saturated.
- $r_{n,l}$ : the number of playing unsaturated suboptimal arms, which is called interruptions during  $I_{n,l}$ .

Let us consider

$$\mathbb{P}[G_n] = \underbrace{\mathbb{P}[G_n, F_{n,K-1}]}_{(D1)} + \underbrace{\mathbb{P}[G_n, F_{n,K-1}^c]}_{(E1)}. \quad (5.31)$$

### Bounds on (D1)

From the definition, one can rewrite

$$(D1) = \mathbb{P}[\{\exists s \in I_{n,K}, \exists a \neq 1 : \tilde{\mu}_a(s) \geq \mu_2 + \epsilon\}, G_n, F_{n,K-1}]$$

$$+ \mathbb{P}[\{\forall s \in I_{n,K}, \forall a \neq 1 : \tilde{\mu}_a(s) \leq \mu_2 + \epsilon\}, G_n, F_{n,K-1}]$$

$$\leq \frac{2(K-1)}{t^3} + \underbrace{\mathbb{P}[\{\forall s \in I_{n,K}, \forall a \neq 1 : \tilde{\mu}_a(s) \leq \mu_2 + \epsilon\}, G_n, F_{n,K-1}]}_{(D2)},$$

$$=: D_{n,K}$$

where the second inequality holds from Lemma 5.16. Here, (D2) can be decomposed as

$$(D2) = \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \exists s \in I_{n,K} : \mathcal{B}_a^c(s) \cup \tilde{\mathcal{B}}_a^c(s)\}]$$

$$+ \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s)\}].$$

From Lemma 5.16, we obtain

$$\begin{aligned}
(D2) &\leq \frac{6(K-1)}{t^3} + \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s)\}] \\
&\leq \frac{6(K-1)}{t^3} \\
&\quad + \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s), \tilde{m}(s) \neq 1\}] \\
&\quad + \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s) \\
&\quad \quad \quad , \{\exists s \in I_{n,K} : \tilde{m}(s) = 1\}\}] \\
&\leq \frac{6(K-1)}{t^3} + C^{\frac{\sqrt{t}-1}{K}} \\
&\quad + \left. \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s)\} \right\} (D3), \\
&\quad \quad \quad , \{\exists s \in I_{n,K} : \tilde{m}(s) = 1\}]
\end{aligned}$$

where the last inequality holds from Lemma 5.17. Next, one can see

$$\begin{aligned}
(D3) &= \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s)\} \\
&\quad \quad \quad , \{\exists s \in I_{n,K} : \tilde{m}(s) = 1, m(s) = 1\}] \\
&\quad + \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s)\} \\
&\quad \quad \quad , \{\exists s \in I_{n,K} : \tilde{m}(s) = 1, m(s) \neq 1\}] \\
&\leq \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s)\} \\
&\quad \quad \quad , \{\exists s \in I_{n,K} : \tilde{m}(s) = 1, m(s) = 1\}] \\
&\quad + \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\text{arm 1 is saturated}\}, \{\exists s \in I_{n,K} : \mathcal{B}_1^c(s)\}] \quad (5.32)
\end{aligned}$$

where (5.32) holds from the sampling procedure since  $i(t) \neq 1$  on  $\mathcal{M}^c(t)$  implies that  $N_1(t) \geq N_{i(t)}$ , i.e., arm 1 is saturated. From Lemma 5.17, it holds that

$$\begin{aligned}
(D3) &\leq \frac{2(K-1)}{t^3} + \mathbb{P}[D_{n,K}, G_n, F_{n,K-1}, \{\forall a \neq 1, \forall s \in I_{n,K} : \mathcal{B}_a(s) \cap \tilde{\mathcal{B}}_a(s)\} \\
&\quad \quad \quad , \{\exists s \in I_{n,K} : \tilde{m}(s) = m(s) = 1\}] \\
&= \frac{2(K-1)}{t^3} + (D4),
\end{aligned}$$

where (D4) denotes the second term. Note that the sampling rule with  $\{m(s) = 1\}$  will choose only  $j(s)$  under the event  $G_n$ , i.e., only  $\{i(s) = j(s)\}$  happens during  $I_n$  for any  $n$  when  $m(s) = \tilde{m}(s)$  holds. It holds that

$$\begin{aligned}
(D4) &\leq \underbrace{\sum_{s \in I_{n,K}} \sum_{a=2}^K \mathbb{P}[m(s) = 1, i(s) = j(s) = a, \mathcal{A}_1(s), \mathcal{B}_a(s), \mathcal{M}(s), G_n]}_{(D5)} \\
&\quad + \underbrace{\sum_{s \in I_{n,K}} \sum_{a=2}^K \mathbb{P}[m(s) = 1, i(s) = j(s) = a, \mathcal{A}_1^c(s), \mathcal{B}_a(s), \mathcal{M}(s)]}_{(D6)}.
\end{aligned}$$

From Lemma 5.7, if an event in (D5) occurs for some  $s$ , then it implies that  $\mathcal{B}_1(t)$  holds for all  $t \geq s$  such that for all  $t \geq N'$ ,  $C_a^* \log t \geq \max\{M, D_1/\underline{d}_a\}$  for all  $a \in [K] \setminus \{1\}$  holds, which contradicts to the event  $G_n$  that implies the existence of  $t \geq s$  such that  $\mathcal{B}_1^c(t)$  holds. Therefore, we have

$$(D5) = 0.$$



Note that (D6) is the form considered in Lemma 5.9. Therefore, we have

$$(D6) \leq \frac{\sqrt{t}-1}{K} \sum_{a=2}^K \mathbb{P}[N_a(s) \leq c_a^* D_a],$$

for some constants  $c_a^*$  and random variables  $D_a$  in Lemma 5.9 such that its expectation is finite. Let  $N_{\mu,A}(\epsilon)$  be a constant that depends on the model and epsilon such that for  $t \geq N_{\mu,A}(\epsilon)$ , it holds for any  $a \in \{2, \dots, K\}$

$$C_a^* \log t \geq c_a^* D_a,$$

i.e., the event in (D6) cannot occur for  $t \geq N_{\mu,A}(\epsilon)$ . Hence, there exist some constant  $C_D(\pi_j, \mu, b, \epsilon) < \infty$  such that

$$\begin{aligned} \sum_{t=1}^T \sum_{n=0}^{\lfloor \sqrt{t} \rfloor} (D1) &\leq \max\{N', N_{\mu,A}(\epsilon)\} + \sum_{t=N_{\mu,A}(\epsilon)+1}^{\infty} \frac{8(K-1)}{t^2 \sqrt{t}} + \sqrt{t} C^{\frac{\sqrt{t}-1}{K}} \\ &\leq C_D(\pi_j, \mu, b, \epsilon). \end{aligned} \quad (5.33)$$

### Bounds on (E1)

By adapting the proof of Kaufmann et al. [2012b], Korda et al. [2013], we prove (E1) is upper bounded by some constants through the mathematical induction, i.e., we will show

$$\mathbb{P}[G_n, F_{n,K-1}^c] \leq (K-2) \left( \frac{10(K-1)}{t^3} + k(\mu, b, n, t) \right),$$

where  $k$  is a function such that  $\sum_{t \geq 1} \sum_{n \leq \sqrt{t}} k < \infty$ .

First, for the base case, it can be easily seen that for  $t \geq N_{\mu,b}$  such that

$$\forall t \geq N_{\mu,b}, \left\lceil \frac{\sqrt{t}-1}{K^2} \right\rceil \geq C_* \log t,$$

where  $C_* = \max_{a \neq 1} C_a$  since only suboptimal arms are selected during  $I_{n,l}$  under  $G_n$ . Then, for  $t \geq N_{\mu,b}$ ,

$$\mathbb{P}[G_n, F_{n,1}^c] = 0.$$

We refer the reader to Kaufmann et al. [2012b] for more explanations in the base case. Then, we assume that for some  $2 \leq l \leq K-1$  if  $t \geq N_{\mu,b}$ , then

$$\mathbb{P}[G_n, F_{n,l-1}^c] \leq (l-2) \left( \frac{10(K-1)}{t^3} + k(\mu, b, n, t) \right).$$

Therefore, we remain to show that

$$\mathbb{P}[G_n, F_{n,l}^c, F_{n,l-1}] \leq \frac{10(K-1)}{t^3} + k(\mu, b, n, t).$$

On the event  $(G_n, F_{n,l}^c, F_{n,l-1})$ , there are exactly  $l-1$  saturated suboptimal arms at the beginning of interval  $I_{n,l}$  and no new arm is saturated during this interval, which implies that  $r_{n,l} \leq KC_* \log t$ . For the set of saturated suboptimal arms  $\mathcal{S}_l$  at the end of  $I_{n,l}$ , it holds that

$$\begin{aligned} \mathbb{P}[G_n, F_{n,l}^c, F_{n,l-1}] &\leq \mathbb{P}[G_n, F_{n,l-1}, \{r_{n,l} \leq KC_* \log t\}] \\ &\leq \mathbb{P}[G_n, F_{n,l-1}, \{\exists s \in I_{n,l}, a \in \mathcal{S}_{l-1} : \tilde{\mathcal{B}}_a^c(s) \cup \mathcal{B}_a^c(s)\}] \\ &\quad + \mathbb{P}[G_n, F_{n,l-1}, \{r_{n,l} \leq KC_* \log t, \\ &\quad \{\forall s \in I_{n,l}, a \in \mathcal{S}_{l-1} : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s)\}\}] \quad (E2), \end{aligned}$$

By applying Lemma 5.16 again, we have

$$\mathbb{P}[G_n, F_{n,l-1}, \{\exists s \in I_{n,l}, a \in \mathcal{S}_{l-1} : \tilde{\mathcal{B}}_a^c(s) \cup \mathcal{B}_a^c(s)\}] \leq \frac{6(K-1)}{t^3}.$$

To bound (E2), we introduce a random interval  $\mathcal{J}_k$  for  $k \in \{0, \dots, r_{n,l} - 1\}$  as the time between  $k$ -th and  $k+1$ -th interruption in  $I_{n,l}$  and set  $\mathcal{J}_k = \emptyset$  for  $k \geq r_{n,l}$ . On (E2), there is a subinterval where no interruptions occur with length  $\lceil \frac{\sqrt{t}-1}{C_* K^2 \log t} \rceil$ . Then, it holds that

$$\begin{aligned} (E2) &\leq \mathbb{P} \left[ \left\{ \exists k \in \{0, \dots, r_{n,l}\} : |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \right. \\ &\quad \left. \left\{ \forall s \in I_{n,l}, a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s) \right\}, G_n, F_{n,l-1} \right] \\ &\leq \sum_{k=1}^{KC_* \log t} \mathbb{P} \left[ \left\{ |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \left\{ \forall s \in \mathcal{J}_k, a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s) \right\}, G_n \right]. \end{aligned}$$

Note that on  $G_n$  and  $\forall s \in \mathcal{J}_k$ , only  $i(s) \in \mathcal{S}_l$  happens, i.e.,  $\{m(s) \neq \tilde{m}(s), m(s) \notin \mathcal{S}_l, \tilde{m}(s) \notin \mathcal{S}_l\}$  cannot occur. Therefore, for any  $s \in \mathcal{J}_k$  under  $\{\forall a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s)\}$ , we have

$$\begin{aligned} \mathbb{1}[m(s) \neq \tilde{m}(s), G_n, \tilde{\mathcal{B}}_{\tilde{m}(s)}(s)] &= \mathbb{1}[m(s) \in \mathcal{S}_l, \tilde{m}(s) \in \mathcal{S}_l \setminus \{m(s)\}, G_n, \tilde{\mathcal{B}}_{\tilde{m}(s)}(s)] \\ &\quad + \mathbb{1}[m(s) = 1, \tilde{m}(s) \in \mathcal{S}_l, G_n, \tilde{\mathcal{B}}_{\tilde{m}(s)}(s), \tilde{\mathcal{B}}_1^c(s)] \\ &\quad + \mathbb{1}[m(s) \in \mathcal{S}_l, \tilde{m}(s) = 1, G_n, \tilde{\mathcal{B}}_1(s), \mathcal{B}_1^c(s)]. \end{aligned}$$

Here, it holds that

$$\{m(s) \in \mathcal{S}_l, \tilde{m}(s) \in \mathcal{S}_l \setminus \{m(s)\}, G_n, \tilde{\mathcal{B}}_{\tilde{m}(s)}(s)\} \subset \{\tilde{\mu}_1(s) \leq \mu_2 + \epsilon, G_n\}.$$

Similarly to the (D3),  $i(s) \neq 1$  implies that arm 1 is already played more than the saturated arm. Let us define an event

$$E2(s) := \{m(s) = \tilde{m}(s) \in \mathcal{S}_l^c \cup \{1\}\} \cap \{\tilde{\mu}_1(s) \geq \mu_2 + \epsilon\}.$$

Then, from the above inclusive relationship, we have

$$\begin{aligned} &\mathbb{P} \left[ \left\{ |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \left\{ \forall s \in \mathcal{J}_k, a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s) \right\}, G_n \right] \\ &\leq \mathbb{P} \left[ \left\{ |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \left\{ \forall s \in \mathcal{J}_k : \left\{ \forall a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s) \right\} \right. \right. \\ &\quad \left. \left. \cap \{\tilde{\mu}_1(s) \leq \mu_2 + \epsilon\} \right\}, G_n \right] \\ &\quad + \mathbb{P} \left[ \left\{ |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \left\{ \forall s \in \mathcal{J}_k, a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s) \right\}, \right. \\ &\quad \left. \left\{ \exists s \in \mathcal{J}_k : \mathcal{B}_1^c(s) \cup \tilde{\mathcal{B}}_1^c(s) \right\}, G_n \right] \\ &\quad + \mathbb{P} \left[ \left\{ |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \left\{ \forall s \in \mathcal{J}_k, a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s) \right\} \right. \\ &\quad \left. \left\{ \exists s \in \mathcal{J}_k : E2(s) \right\}, G_n \right] \Big\}^{(E3)}. \end{aligned}$$

By applying Lemmas 5.16 and 5.17, we have

$$\mathbb{P} \left[ \left\{ |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \{\forall s \in \mathcal{J}_k, a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s)\}, G_n \right] \leq C \frac{\sqrt{t}-1}{C_* K^2 \log t} + \frac{6}{t^3} + (E3).$$

From the definition of  $\mathcal{J}_k$  and  $G_n$ , one can see that

$$\begin{aligned} (E3) &= \mathbb{P} \left[ \left\{ |\mathcal{J}_k| \geq \frac{\sqrt{t}-1}{C_* K^2 \log t} \right\}, \{\forall s \in \mathcal{J}_k : a \in \mathcal{S}_l : \tilde{\mathcal{B}}_a(s) \cap \mathcal{B}_a(s)\} \right. \\ &\quad \left. , \{\exists s \in \mathcal{J}_k : E2(s) \cap \{j(s) = i(s) \in \mathcal{S}_l\}\}, G_n \right] \\ &\leq \mathbb{P} \left[ \exists s \in \mathcal{J}_k : m(s) = \tilde{m}(s) \in \mathcal{S}_l^c \cup \{1\}, j(s) \in \mathcal{S}_l, i(s) = j(s), \mathcal{A}_{m(s)}^c \right. \\ &\quad \left. , \mathcal{B}_{j(s)}, \tilde{\mu}_1(s) \geq \mu_2 + \epsilon, G_n \right] \\ &\quad + \mathbb{P} \left[ \exists s \in \mathcal{J}_k : m(s) = \tilde{m}(s) \in \mathcal{S}_l^c \cup \{1\}, j(s) \in \mathcal{S}_l, i(s) = j(s), \mathcal{A}_{m(s)} \right. \\ &\quad \left. , \mathcal{B}_{j(s)}, \tilde{\mu}_1(s) \geq \mu_2 + \epsilon, G_n \right]. \quad (5.34) \\ &=: (E4) + (E5). \end{aligned}$$

The first equation holds since only saturated suboptimal arms have to be played on  $\mathcal{J}_k$  when  $m(s) = \tilde{m}(s)$  is unsaturated or optimal arm, which makes  $j(s) = i(s) \in \mathcal{S}_l$ . Let us denote the event in the first term and the second term of RHS in (5.34) by (E4) and (E5), respectively.

From Lemma 5.9, we have

$$\begin{aligned} \mathbb{1}[(E4)] &\leq \sum_{s \in \mathcal{J}_k} \sum_{a \in \mathcal{S}_l} \sum_{m \in \mathcal{S}_l \cup \{1\}} \mathbb{1}[m(s) = m, i(s) = j(s) = a, \mathcal{A}_m^c(s), \mathcal{B}_a(s)] \\ &\leq \sum_{s \in \mathcal{J}_k} \sum_{a \in \mathcal{S}_l} \sum_{m \in \mathcal{S}_l \cup \{1\}} \mathbb{1}[N_a(s) \leq c_{m,a}^* D_{m,a}]. \end{aligned}$$

Similarly to the case of (D4), there exists some deterministic constant  $N_{\mu,A}(\epsilon)'$  such that for  $t \geq N_{\mu,A}(\epsilon)'$ ,  $\forall (m, a) \in (\mathcal{S}_l^c \cup \{1\}, \mathcal{S}_l)$

$$C_a^* \log t \geq c_{m,a}^* D_{m,a},$$

where we replace 1 by  $m$  in  $c_a^*$  and  $D_a$  to define  $c_{m,a}^*$  and  $D_{m,a}$ .

Further, (E5) can be decomposed by

$$(E5) = (E6) + (E7),$$

where

$$\begin{aligned} (E6) &:= \mathbb{P} \left[ \exists s \in \mathcal{J}_k : m(s) = \tilde{m}(s) \in \mathcal{S}_l^c, j(s) \in \mathcal{S}_l, i(s) = j(s), \right. \\ &\quad \left. \mathcal{A}_{m(s)}, \mathcal{B}_{j(s)}, \tilde{\mu}_1(s) \geq \mu_2 + \epsilon, G_n \right] \\ (E7) &:= \mathbb{P} \left[ \exists s \in \mathcal{J}_k : m(s) = \tilde{m}(s) = 1, j(s) \in \mathcal{S}_l, i(s) = j(s), \right. \\ &\quad \left. \mathcal{A}_1, \mathcal{B}_{j(s)}, \tilde{\mu}_1(s) \geq \mu_2 + \epsilon, G_n \right]. \end{aligned}$$

Note that on  $(E6)$ ,  $\tilde{\mathcal{B}}_m^c(s)$  always holds since  $\tilde{\mu}_1 > \mu_2 + \epsilon$  but  $\tilde{m}(s) \neq 1$  and  $(E5)$  is a subset of the event we consider in Lemma 5.7, i.e., event  $(E6)$  implies the existence of  $s \in \mathcal{J}_k$  such that

$$N_m(s) \geq N_{j(s)} \frac{d_{j(s)}}{d(\mu_m + \epsilon, \mu_j - \epsilon)} \geq C_* \frac{d_{j(s)}}{d(\mu_m + \epsilon, \mu_{j(s)} - \epsilon)} \log t.$$

From the definition of  $C_*$  and saturation, it holds that for any  $m \in \mathcal{S}_l^c$

$$C_* \frac{d_{j(s)}}{d(\mu_m + \epsilon, \mu_{j(s)} - \epsilon)} \geq C_* \frac{\min_{a \neq 1} d_a}{d(\mu_2 + \epsilon, \mu_K - \epsilon)} \geq C'_m \log t.$$

As a result, we have

$$\mathbb{P}[(E6)] = \mathbb{P}[\{\exists s \in \mathcal{J}_k, m \in \mathcal{S}_l^c : \tilde{\mathcal{B}}_m^c(s)\} \cap (E5)] \leq \frac{4(K-1)}{t^3}.$$

Similarly to the case of  $(D5)$ , if the event in  $(E7)$  occurs some  $s \in \mathcal{J}_k$  for  $t$  such that  $t \geq N'$ ,  $C_a^* \log t \geq \max\{M, D_1/d_a\}$  for all  $a \in [K] \setminus \{1\}$ , then only  $\mathcal{B}_1(t)$  holds for  $s \geq t$  holds, which contradicts to the event  $G_n$ .

Therefore, for  $t \geq N_0 := \max(N_{\mu, b}, N_{\mu, A}(\epsilon)', N_K, N')$ , where  $N_K$  in Lemma 5.16, it holds

$$(E2) \leq KC_* \log t \left( C^{\frac{\sqrt{t}-1}{C_* K^2 \log t}} + \frac{10(K-1)}{t^3} \right) =: k(\mu, b, n, t).$$

Hence, there exists some constants  $C_E(\pi_j, \mu, b, \epsilon) < \infty$  such that

$$\begin{aligned} \sum_{T=1}^{\infty} \sum_{t=T+1}^{\infty} \sum_{n=1}^{\lfloor \sqrt{t} \rfloor} (E1) &\leq N_0 + \sum_{T=N_0+1}^{\infty} \sum_{t=T+1}^{\infty} \frac{6(K-1)^2}{t^2 \sqrt{t}} \\ &\quad + \sum_{T=N_0+1}^{\infty} \sum_{t=T+1}^{\infty} KC_* \log t \left( \sqrt{t} C^{\frac{\sqrt{t}-1}{C_* K^2 \log t}} + \frac{10(K-1)}{t^2 \sqrt{t}} \right) \\ &\leq N_0 + C_E(\pi_j, \mu, b, \epsilon). \end{aligned} \tag{5.35}$$

### (3) Conclusion

By combining (5.33) and (5.35) with (5.31), we obtain

$$\begin{aligned} \sum_{T=1}^{\infty} \sum_{t=T+1}^{\infty} \mathbb{P}[N_1(t) \leq \sqrt{t}, \mathcal{C}(t)] &\leq \sum_{T=1}^{\infty} \sum_{t=T+1}^{\infty} \sum_{n=N_1(T+1)}^{\lfloor \sqrt{t} \rfloor} (D1) + (E1) \\ &\leq N_0 + C_D(\pi_j, \mu, b, \epsilon) + C_E(\pi_j, \mu, b, \epsilon) \\ &=: C(\pi_j, \mu, b, \epsilon) < \infty, \end{aligned}$$

which concludes the proof.  $\square$

### 5.6.9 Proof of Theorem 5.4: Sample complexity

Here, we derive the upper bound on the sample complexity of BC-TE.

Before beginning the proof, we first provide a technical lemma provided in Garivier and Kaufmann [2016].

**Lemma 5.18** (Lemma 18 in Garivier and Kaufmann [2016]). *For every  $\alpha \in [1, \frac{e}{2}]$ , for any two constants  $c_1, c_2 > 0$ ,*

$$x = \frac{\alpha}{c_1} \left[ \log \left( \frac{c_2 e}{c_1^\alpha} \right) + \log \log \left( \frac{c_2}{c_1^\alpha} \right) \right]$$

*is such that  $c_1 x \geq \log(c_2 x^\alpha)$ .*

Next, we define a set of bandit instances  $\mathcal{S}$  for any  $\epsilon > 0$  as follows:

$$\mathcal{S} = \mathcal{S}(\nu, \epsilon) := \{\boldsymbol{\mu}' : |\boldsymbol{\mu}' - \boldsymbol{\mu}| \leq \epsilon\},$$

where  $\boldsymbol{\mu}$  denotes the true mean reward vector. For any  $i \neq 1$ , if  $\boldsymbol{\mu}' \in \mathcal{S}$ , we have the following inequality:

$$\forall \mathbf{w} \in \Sigma_K : \frac{1}{1+\epsilon} f_i(\mathbf{w}; \boldsymbol{\mu}) \leq f_i(\mathbf{w}; \boldsymbol{\mu}') \leq (1+\epsilon) f_i(\mathbf{w}; \boldsymbol{\mu}). \quad (5.36)$$

From the relationship in (5.14), (5.36) is equivalent to

$$\begin{aligned} \forall \mathbf{w} \in \Sigma_K : \frac{1}{1+\epsilon} g(\mathbf{w}; \boldsymbol{\mu}) &\leq g(\mathbf{w}; \boldsymbol{\mu}') \leq (1+\epsilon) g(\mathbf{w}; \boldsymbol{\mu}) \\ \forall x \in [0, 1] : \frac{1}{1+\epsilon} k_i(x; \boldsymbol{\mu}) &\leq k_i(x; \boldsymbol{\mu}') \leq (1+\epsilon) k_i(x; \boldsymbol{\mu}) \\ \forall z \in [0, 1] : \frac{1}{1+\epsilon} h_i(z; \boldsymbol{\mu}) &\leq h_i(z; \boldsymbol{\mu}') \leq (1+\epsilon) h_i(z; \boldsymbol{\mu}). \end{aligned}$$

Notice that that for any  $t \geq T_B$ ,  $\hat{\boldsymbol{\mu}}(t) \in \mathcal{S}$  holds from the the definition of  $T_B$  in (5.8).

Therefore, we can assume

$$\frac{1}{1+\epsilon} \frac{z_i^*}{1-z_i^*} \leq \frac{z_i^*(\boldsymbol{\mu}')}{1-z_i^*(\boldsymbol{\mu}')} \leq (1+\epsilon) \frac{z_i^*}{1-z_i^*} \quad (5.37)$$

$$\frac{1}{1+\epsilon} \frac{z_i}{1-z_i} \leq \frac{z_i(\boldsymbol{\mu}')}{1-z_i(\boldsymbol{\mu}')} \leq (1+\epsilon) \frac{z_i}{1-z_i}. \quad (5.38)$$

and for  $t \geq T_B$  and the definition of a challenger at round  $t$ ,  $j(t)$  in (5.7),

$$\frac{1}{1+\epsilon} \min_{a \neq 1} f_i(x; \boldsymbol{\mu}) \leq f_{j(t)}(x; \boldsymbol{\mu}) \leq (1+\epsilon) \min_{a \neq 1} f_i(x; \boldsymbol{\mu}). \quad (5.39)$$

Notice that (5.39) provides

$$\frac{1}{1+\epsilon} \min_{a \neq 1} k_i(x; \boldsymbol{\mu}) \leq k_{j(t)}(x; \boldsymbol{\mu}) \leq (1+\epsilon) \min_{i \neq 1} k_i(x; \boldsymbol{\mu}). \quad (5.40)$$

Since  $t f_i(\mathbf{w}^t; \boldsymbol{\mu}) = (N_1(t) + N_i(t)) h_i(z_i^t; \boldsymbol{\mu})$  holds from their relationship in (5.14) and  $z_i^t = \frac{w_i^t}{w_1^t + w_i^t}$ , (5.39) also implies that

$$\begin{aligned} \frac{1}{1+\epsilon} \min_{i \neq 1} (N_1(t) + N_i(t)) h_i(z_i^t; \boldsymbol{\mu}) &\leq (N_1(t) + N_{j(t)}(t)) h_{j(t)}(z_{j(t)}^t; \boldsymbol{\mu}) \\ &\leq (1+\epsilon) \min_{i \neq 1} (N_1(t) + N_i(t)) h_i(z_i^t; \boldsymbol{\mu}). \end{aligned}$$

From the concavity of the objective function, we have the following result, whose proof is provided in Section 5.6.10.

**Lemma 5.19.** *For any  $i \neq 1$ ,  $t f_i(\mathbf{w}^t; \boldsymbol{\mu})$  is non-decreasing with respect to  $t \in \mathbb{N}$ .*

*Proof of theorem 5.4.* We first introduce a positive increasing sequence  $(G_m)_{m \in \mathbb{N}}$  and let  $\psi_m$  be the first round where  $tg(\mathbf{w}^t; \boldsymbol{\mu}) > G_m$  holds, which is defined as

$$\psi_m := \inf\{t \in \mathbb{N}_{\geq T_B} : tg(\mathbf{w}^t; \boldsymbol{\mu}) \geq G_m\}.$$

Notice that Lemma 5.19 ensures  $\psi_m \leq \psi_{m+1}$  for any  $m \in \mathbb{N}$  since  $tg(\mathbf{w}^t; \boldsymbol{\mu}) = t \min_{i \neq 1} f_i(\mathbf{w}^t; \boldsymbol{\mu})$  is non-decreasing.

For notational simplicity,  $\underline{g}$  denotes the value of the objective function  $g(\mathbf{w}; \boldsymbol{\mu})$  at  $\mathbf{w} = \underline{\mathbf{w}}$  defined in (5.16). Then from (5.14)

$$\forall i \neq 1 : \underline{g} = \underline{w}_1 k_i(\underline{w}_i / \underline{w}_1; \boldsymbol{\mu}) = (\underline{w}_1 + \underline{w}_i) h_i(\underline{z}_i; \boldsymbol{\mu}). \quad (5.41)$$

Here, we set  $G_1$  to satisfy

$$\forall i \in [K] : N_i(T_B) \leq \frac{\underline{w}_i}{\underline{g}} G_1. \quad (5.42)$$

Then, the stopping time  $\tau_\delta$  can be written as

$$\begin{aligned} \tau_\delta &= \inf\{t \in \mathbb{N} : tg(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t)) \geq \beta(t, \delta)\} \\ &\leq \inf\{t \in \mathbb{N}_{\geq T_B} : \frac{tg(\mathbf{w}^t; \boldsymbol{\mu})}{1 + \epsilon} \geq \beta(t, \delta)\} \\ &\leq T_B + \inf\left\{\psi_m : \frac{1}{1 + \epsilon} G_m \geq \beta(\psi_m, \delta), m \in \mathbb{N}\right\}. \end{aligned} \quad (5.43)$$

To find the upper bound of the stopping time, we require the relationship between  $G_m$  and  $\psi_m$ . To do this, we first derive the bounds on the number of plays  $N_i(t)$ .

### Bounds on the number of plays

Here, we aim to derive the upper bounds on  $N_i(t)$  for  $t \in [\psi_m, \psi_{m+1})$  and for any  $i \in [K]$ .

For  $t \geq T_B$ , only  $m(t) = 1$  occurs. Therefore, an arm  $i \neq 1$  is played either when TE occurs or when  $j(t) = i$  and  $d(\hat{\mu}_i(t), \hat{\mu}_{1,i}(t)) \geq d(\hat{\mu}_1(t), \hat{\mu}_{1,i}(t))$  for  $t \geq T_B$ . Thus, if  $j(t) \neq i$  holds for all  $t \in [\psi_m, \psi_{m+1})$ , then

$$N_i(\psi_{m+1}) = N_i(\psi_m) + M_{i,m},$$

where  $M_{i,m}$  denote the number of the arm  $i$  being played by TE during  $[\psi_m, \psi_{m+1})$ , which is

$$M_{i,m} = \sum_{t=\psi_m}^{\psi_{m+1}-1} \mathbb{1}[\mathcal{M}^c(t), i(t) = i].$$

The latter condition can be rewritten as  $j(t) = i$  and  $z_i^t \leq z_i^*(\hat{\boldsymbol{\mu}}(t))$  from the definition of  $z_i^*$  in (5.13). For notational simplicity, we denote  $z_i^*(\hat{\boldsymbol{\mu}}(t))$  and  $\underline{z}_i(\hat{\boldsymbol{\mu}}(t))$  by  $z_{i,t}^*$  and  $\underline{z}_{i,t}$ , respectively.

**(1) Upper bound for the second-best arm** Firstly, let us consider the second-best arm  $j^*(\nu)$ , which is assumed to be the arm 2 in this chapter. It should be noted that the second-best arm may not be unique. Then let us define a partition of  $Q_m := [\psi_m, \psi_{m+1})$

$$\begin{aligned} (Q1) &:= \left\{t \in [\psi_m, \psi_{m+1}) : N_1(t) \leq \frac{\underline{w}_1}{\underline{g}} G_{m+1}\right\} \\ (Q2) &:= \left\{t \in [\psi_m, \psi_{m+1}) : N_1(t) > \frac{\underline{w}_1}{\underline{g}} G_{m+1}\right\}. \end{aligned}$$

Then, we define  $\epsilon_1 = \epsilon_1(\epsilon, G_{m+1}/G_m) > \epsilon$  to be a constant satisfying

$$k_2 \left( (1 + \epsilon_1) \frac{w_2}{w_1}; \boldsymbol{\mu} \right) \geq \frac{G_{m+1}}{G_m} \frac{g}{w_1}, \quad (5.44)$$

Here, one can see that  $\epsilon_1 \rightarrow 0_+$  as  $\epsilon \rightarrow 0_+$  and  $\frac{G_{m+1}}{G_m} \rightarrow 1_+$  from (5.41). Then we will show that if  $N_2(t) \geq N' = (1 + \epsilon_1) \frac{w_2}{g} G_{m+1}$ , then  $i(t) = 2$  holds only when TE occurs.

**(1-i) When  $t \in (Q1)$**  In this case,

$$\begin{aligned} N_2(t) &\geq N' = (1 + \epsilon_1) \frac{w_2}{g} G_m = (1 + \epsilon_1) \frac{w_2}{w_1} \frac{w_1}{g} G_m \\ &\geq (1 + \epsilon_1) \frac{w_2}{w_1} N_1(t) \quad \because t \in (Q1) \\ &= (1 + \epsilon_1) \frac{z_2}{1 - z_2} N_1(t) \quad \text{by definition of } \underline{w} \text{ in (5.16)} \\ &= (1 + \epsilon_1) \frac{z_2^*}{1 - z_2^*} N_1(t) \quad \text{by definition of } \underline{z} \text{ in (5.15)} \\ &> \frac{z_{2,t}^*}{1 - z_{2,t}^*} N_1(t). \quad \text{by (5.37) and } \epsilon_1 > \epsilon \end{aligned}$$

This implies that for  $t \in (Q1)$ , if  $N_2(t) \geq N'$ , then  $z_2^t > z_{2,t}^*$  holds. Therefore, only  $i(t) = 1$  happens unless TE occurs.

**(1-ii) When  $t \in (Q2)$**  From the relationship between  $f_i$  and  $k_i$  in (5.14), one can see that  $tf_i(w^t; \boldsymbol{\mu}) = N_1(t)k_i(w_i^t/w_1^t; \boldsymbol{\mu})$ . Therefore, one can extend Lemma 5.19 to show that  $yk_i(c/y; \boldsymbol{\mu})$  is non-decreasing with respect to  $y \geq 0$  for fixed  $c > 0$  and any  $i \neq 1$ . Recall that the  $k_i(x; \boldsymbol{\mu})$  is a strictly increasing function with respect to  $x > 0$ . Then we can obtain that

$$\begin{aligned} N_1(t)k_2 \left( \frac{N_2(t)}{N_1(t)}; \boldsymbol{\mu} \right) &\geq N_1(t)k_2 \left( \frac{N'}{N_1(t)}; \boldsymbol{\mu} \right) \\ &\geq G_m \frac{w_1}{g} k_2 \left( N' \frac{g}{G_m w_1}; \boldsymbol{\mu} \right) \quad \because t \in (Q2) \\ &= G_m \frac{w_1}{g} k_2 \left( (1 + \epsilon_1) \frac{w_2}{w_1}; \boldsymbol{\mu} \right) \\ &\geq G_m \frac{w_1}{g} \frac{G_{m+1}}{G_m} \frac{g}{w_1} \quad \text{by definition of } \epsilon_1 \text{ in (5.44)} \\ &= G_{m+1}, \end{aligned}$$

which contradicts the assumption  $t \in (Q2)$ .

**(1-iii) Conclusion** Therefore, for any  $t \in Q_m$ ,

$$\left\{ N_2(t) \geq (1 + \epsilon_1) \frac{w_2}{g} G_m \right\} \implies \{j(t) \neq 2\},$$

which directly implies that

$$N_2(t) \leq \max \left( N_2(\psi_m), (1 + \epsilon_1) \frac{w_2}{g} G_m \right) + M_{2,m}.$$

Here, from the definition of  $G_1$  in (5.42),  $N_1(t) \leq \frac{w_1}{\underline{g}} G_1$  holds for all  $t < \psi_1$ , which implies that  $N_2(\psi_m) \leq (1 + \epsilon_1) \frac{w_2}{\underline{g}} G_m + M_{2,0}$ . Therefore, for any  $t \in [\psi_m, \psi_{m+1})$ ,

$$N_2(t) \leq (1 + \epsilon_1) \frac{w_2}{\underline{g}} G_m + M_2(\psi_{m+1})$$

where  $M_i(\psi_{m+1}) = \sum_{l=0}^m M_{i,l}$  for any  $i \in [K]$ .

Here, let us define a random variable  $M_T = \sum_{t=T_B}^T \mathbb{1}[\mathcal{M}^c(t)] = \sum_{i=1}^K \sum_m M_{i,m}$ , which satisfies  $\mathbb{E}[M_T] < \infty$  by Lemma 5.11. Then we can set  $G_m$  sufficiently large to satisfy

$$G_m \geq \frac{g}{\epsilon} M_T,$$

which directly implies that

$$N_2(t) \leq (1 + \epsilon_1) \frac{w_2}{\underline{g}} G_m + \frac{\epsilon}{\underline{g}} G_m. \quad (5.45)$$

**(2) Lower bound for the optimal arm** For any  $t \in Q_m$ , it holds that

$$\begin{aligned} G_m &\leq N_1(t) \min_{i \neq 1} k_i \left( \frac{N_i(t)}{N_1(t)}; \boldsymbol{\mu} \right) \\ &= \min_{i \neq 1} (N_1(t) + N_i(t)) h_i(z_i^t; \boldsymbol{\mu}) && \text{by (5.14)} \\ &\leq (N_1(t) + N_2(t)) h_2(z_2^t; \boldsymbol{\mu}) \\ &\leq (N_1(t) + N_2(t)) h_2(z_2; \boldsymbol{\mu}) && \text{by } z_2 = z_2^* \\ &= \frac{N_1(t) + N_2(t)}{w_1 + w_2} \underline{g}. && \text{by (5.41)} \end{aligned}$$

Therefore, for  $t = \psi_m$ , the upper bound of  $N_2(\psi_m)$  in (5.45) provides

$$N_1(\psi_m) \geq \frac{w_1 + w_2}{\underline{g}} G_m - (1 + \epsilon_1) \frac{w_2}{\underline{g}} G_m - \frac{\epsilon}{\underline{g}} G_m.$$

Since  $N_1(t)$  is non-decreasing from its definition, for any  $t \geq \psi_m$ ,

$$N_1(t) \geq \frac{w_1}{\underline{g}} G_m - \epsilon_1 \frac{w_2}{\underline{g}} G_m - \frac{\epsilon}{\underline{g}} G_m. \quad (5.46)$$

**(3) Upper bound on the challenger arms** Based on the results obtained in (1) and (2), we will derive the upper bound of  $N_{j(t)}(t)$  for  $t \geq T_B$ . For  $t \in Q_m$ , it holds that

$$G_m \leq N_1(t) \min_{i \neq 1} k_i \left( \frac{N_i(t)}{N_1(t)}; \boldsymbol{\mu} \right) < G_{m+1}.$$

Since  $j(t) = \arg \min_{i=1} f_i(\mathbf{w}^t; \hat{\boldsymbol{\mu}}(t))$ , by using (5.40), one can obtain that

$$\frac{1}{1 + \epsilon} k_{j(t)} \left( \frac{N_{j(t)}(t)}{N_1(t)}; \boldsymbol{\mu} \right) \leq \min_{i \neq 1} k_i \left( \frac{N_i(t)}{N_1(t)}; \boldsymbol{\mu} \right).$$

Then, by (5.46)

$$\begin{aligned} N_1(t) \min_{i \neq 1} k_i \left( \frac{N_i(t)}{N_1(t)}; \boldsymbol{\mu} \right) &\geq \frac{1}{1 + \epsilon} N_1(t) k_{j(t)} \left( \frac{N_{j(t)}(t)}{N_1(t)}; \boldsymbol{\mu} \right) \\ &\geq \frac{1}{1 + \epsilon} \frac{G_m}{\underline{g}} (w_1 - \epsilon_1 w_2 - \epsilon) k_{j(t)} \left( \frac{\underline{g} N_{j(t)}(t)}{(w_1 - \epsilon_1 w_2 - \epsilon) G_m}; \boldsymbol{\mu} \right), \end{aligned}$$



which implies

$$k_{j(t)} \left( \frac{\underline{g} N_{j(t)}(t)}{(\underline{w}_1 - \epsilon_1 \underline{w}_2 - \epsilon) G_m}; \boldsymbol{\mu} \right) < (1 + \epsilon) \frac{G_{m+1}}{G_m} \frac{\underline{g}}{\underline{w}_1 - \epsilon_1 \underline{w}_2 - \epsilon}.$$

This directly implies that

$$\begin{aligned} \frac{\underline{g} N_{j(t)}(t)}{(\underline{w}_1 - \epsilon_1 \underline{w}_2 - \epsilon) G_m} &< l_{j(t)} \left( (1 + \epsilon) \frac{G_{m+1}}{G_m} \frac{\underline{g}}{\underline{w}_1 - \epsilon_1 \underline{w}_2 - \epsilon}; \boldsymbol{\mu} \right) \\ &\leq (1 + \epsilon_2) \frac{\underline{w}_{j(t)}}{\underline{w}_1}, \end{aligned}$$

where  $l_i$  is the inverse function of  $k_i$  defined in (5.12) and  $\epsilon_2 > \epsilon_1$  is a constant such that  $\epsilon_2 \rightarrow 0_+$  as  $\epsilon \rightarrow 0_+$  and  $\frac{G_{m+1}}{G_m} \rightarrow 1_+$ . Then, we have for any  $t \in Q_m$  that

$$N_{j(t)}(t) < (1 + \epsilon_2) \frac{\underline{w}_{j(t)}}{\underline{g}} G_m.$$

In other words, if there exist  $s \in Q_m$  such that

$$N_i(t) \geq (1 + \epsilon_2) \frac{\underline{w}_i}{\underline{g}} G_m,$$

then only  $j(s) \neq 1$  occurs for  $t \in [s, \psi_{m+1})$ , which implies that such arm  $i$  will be played only when TE occurs until  $\psi_{m+1}$ . Therefore, for  $t \in Q_m$

$$\begin{aligned} N_i(t) &\leq \max \left( N_i(\psi_m, (1 + \epsilon_2) \frac{\underline{w}_i}{\underline{g}} G_m) \right) + M_{i,m} \\ &\leq (1 + \epsilon_2) \frac{\underline{w}_i}{\underline{g}} G_m + M_i(\psi_{m+1}) \\ &\leq (1 + \epsilon_2) \frac{\underline{w}_i}{\underline{g}} G_m + \frac{\epsilon}{\underline{g}} G_m. \end{aligned}$$

**(4) Upper bound on the optimal arm** Here, let us assume that there exists  $t' \in Q_m$  such that  $N_1(t') \geq (1 + \epsilon)(1 + \epsilon_2) \frac{\underline{w}_1}{\underline{g}} G_m$ . If there exists no such  $t'$ , then one can directly obtain that  $N_1(t) \leq (1 + \epsilon)(1 + \epsilon_2) \frac{\underline{w}_1}{\underline{g}} G_m$  for all  $t \in Q_m$ .

Since  $N_{j(t)}(t) < (1 + \epsilon_2) \frac{\underline{w}_{j(t)}}{\underline{g}} G_m$  holds from (5.6.9), then for any  $t \in [t', \psi_{m+1})$

$$\begin{aligned} \frac{N_{j(t)}(t)}{N_1(t)} &< \frac{1}{1 + \epsilon} \frac{\underline{w}_{j(t)}}{\underline{w}_1} = \frac{1}{1 + \epsilon} \frac{\underline{z}_{j(t)}}{1 - \underline{z}_{j(t)}} \\ &\leq \frac{\underline{z}_{j(t),t}}{1 - \underline{z}_{j(t),t}}, \end{aligned} \quad \text{by (5.38)}$$

which implies that  $\underline{z}_{j(t)}^t < \underline{z}_{j(t),t} \leq \underline{z}_{j(t),t}^*$ . Since BC-TE plays the optimal arm 1 if  $\underline{z}_{j(t),t} \geq \underline{z}_{j(t),t}^*$ , only  $i(t) = j(t)$  is possible unless TE occurs until  $\psi_{m+1}$ . Therefore, for  $t \in Q_m$ , it holds that

$$\begin{aligned} N_1(t) &\leq \max \left( N_1(\psi_m), (1 + \epsilon)(1 + \epsilon_2) \frac{\underline{w}_1}{\underline{g}} G_m \right) + M_{1,m} \\ &\leq (1 + \epsilon)(1 + \epsilon_2) \frac{\underline{w}_1}{\underline{g}} G_m + M_1(\psi_{m+1}) \\ &\leq (1 + \epsilon_3) \frac{\underline{w}_1}{\underline{g}} G_m + \frac{\epsilon}{\underline{g}} G_m, \end{aligned}$$

where  $\epsilon_3$  is a constant such that  $(1 + \epsilon)(1 + \epsilon_2) = 1 + \epsilon_3$ . One can see that  $\epsilon_3 \rightarrow 0_+$  as  $\epsilon \rightarrow 0_+$  and  $\frac{G_{m+1}}{G_m} \rightarrow 1_+$ .

**(5) Conclusion** In summary, for any  $t \in [\psi_m, \psi_{m+1})$ , the results in (1)–(4) imply that for any  $i \in [K]$ :

$$N_i(t) \leq (1 + \epsilon_3) \frac{w_i}{\underline{g}} G_m + \frac{\epsilon}{\underline{g}} G_m. \quad (5.47)$$

### Sample complexity

From the upper bound on the number of plays for each arm in (5.47), for any  $m \in \mathbb{N}$ ,

$$\begin{aligned} \psi_m &= \sum_{i=1}^K N_i(\psi_m) \leq \sum_{i=1}^K (1 + \epsilon_3) \frac{w_i}{\underline{g}} G_m + \frac{\epsilon}{\underline{g}} G_m \\ &= (1 + \epsilon_3) \frac{1}{\underline{g}} G_m + \frac{K\epsilon}{\underline{g}} G_m, \end{aligned}$$

which implies that

$$\frac{\underline{g}\psi_m}{(1 + \epsilon_3 + K\epsilon)} \leq G_m.$$

Therefore, the stopping time  $\tau_\delta$  in (5.43) can be written as

$$\begin{aligned} \tau_\delta &\leq T_B + \inf \left\{ \psi_m : \frac{1}{1 + \epsilon} G_m \geq \beta(\psi_m, \delta) \right\} \\ &\leq T_B + \inf \left\{ \psi_m : \frac{1}{1 + \epsilon} \frac{\underline{g}\psi_m}{(1 + \epsilon_3 + K\epsilon)} \geq \beta(\psi_m, \delta) \right\} \\ &\leq T_B + \inf \left\{ \psi_m : \frac{\underline{g}\psi_m}{(1 + \epsilon_4)} \geq \log \left( \frac{Ct^\alpha}{\delta} \right) \right\}, \end{aligned}$$

for some  $\epsilon_4 > \epsilon_3$  satisfying  $\epsilon_4 \rightarrow 0_+$  as  $\epsilon \rightarrow 0_+$  and  $\frac{G_{m+1}}{G_m} \rightarrow 1_+$  and constants  $C$  and  $\alpha \in [1, e/2]$  considered in Section 5.2.2. Then, by Lemma 5.18

$$\tau_\delta \leq T_B + \frac{\alpha}{\underline{g}} (1 + \epsilon_4) \left[ \log \left( (1 + \epsilon_4)^\alpha \frac{Ce}{\delta \underline{g}^\alpha} \right) + \log \log \left( (1 + \epsilon_4)^\alpha \frac{C}{\delta \underline{g}^\alpha} \right) \right].$$

Therefore, by taking expectations, we can obtain that

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \frac{\alpha(1 + \epsilon_4)}{\underline{g}}$$

since  $\mathbb{E}[T_B]$  is finite from Theorem 5.3. Letting  $\epsilon \rightarrow 0$  and setting  $\frac{G_{m+1}}{G_m} \rightarrow 1$  conclude the proof.  $\square$

### 5.6.10 Proof of Lemma 5.19

Here, we prove Lemma 5.19.

*Proof.* From the relation with  $f_i$  and  $h_i$  in (5.14), we can rewrite the function  $tf_i(\mathbf{w}^t; \boldsymbol{\mu})$  as

$$tf_i(\mathbf{w}^t; \boldsymbol{\mu}) = (N_1(t) + N_i(t)) h_i \left( \frac{N_i(t)}{N_1(t) + N_i(t)}; \boldsymbol{\mu} \right).$$

Recall that  $h_i(z; \boldsymbol{\mu})$  is a concave function with respect to  $z \in [0, 1]$  and  $h_i(0; \boldsymbol{\mu}) = h_i(1; \boldsymbol{\mu}) = 0$  for any  $i \neq 1$ . For any  $i \neq 1$ , let us consider three possible cases (1)  $i(t) = 1$ , (2)  $i(t) = i$ , and (3)  $i(t) \notin \{1, i\}$ .

**(1) When the optimal arm is played** When  $i(t) = 1$  holds, for any  $i \neq 1$

$$(t+1)f_i(\mathbf{w}^{t+1}; \boldsymbol{\mu}) = (N_1(t) + N_i(t) + 1)h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t) + 1}; \boldsymbol{\mu}\right).$$

From the concavity of  $h_i$ , we obtain that

$$\begin{aligned} h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t) + 1}; \boldsymbol{\mu}\right) &= h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t)} \frac{N_1(t) + N_i(t)}{N_1(t) + N_i(t) + 1}; \boldsymbol{\mu}\right) \\ &\geq \frac{N_1(t) + N_i(t)}{N_1(t) + N_i(t) + 1} h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t)}; \boldsymbol{\mu}\right) \\ &\quad + \frac{1}{N_1(t) + N_i(t) + 1} h_i(0; \boldsymbol{\mu}), \end{aligned}$$

which implies

$$\begin{aligned} (N_1(t) + N_i(t) + 1)h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t) + 1}; \boldsymbol{\mu}\right) \\ \geq (N_1(t) + N_i(t))h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t)}; \boldsymbol{\mu}\right) = tf_i(\mathbf{w}^t; \boldsymbol{\mu}). \end{aligned}$$

This concludes the case when  $i(t) = 1$ .

**(2) When the suboptimal arm is played** When  $i(t) = i$  holds,

$$(t+1)f_i(\mathbf{w}^{t+1}; \boldsymbol{\mu}) = (N_1(t) + N_i(t) + 1)h_i\left(\frac{N_i(t) + 1}{N_1(t) + N_i(t) + 1}; \boldsymbol{\mu}\right).$$

By the concavity, again, we obtain that

$$\begin{aligned} h_i\left(\frac{N_i(t) + 1}{N_1(t) + N_i(t) + 1}; \boldsymbol{\mu}\right) \\ = h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t)} \frac{N_1(t) + N_i(t)}{N_1(t) + N_i(t) + 1} + \frac{1}{N_1(t) + N_i(t) + 1}; \boldsymbol{\mu}\right) \\ \geq \frac{N_1(t) + N_i(t)}{N_1(t) + N_i(t) + 1} h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t)}; \boldsymbol{\mu}\right) + \frac{1}{N_1(t) + N_i(t) + 1} h_i(1; \boldsymbol{\mu}) \\ = \frac{N_1(t) + N_i(t)}{N_1(t) + N_i(t) + 1} h_i\left(\frac{N_i(t)}{N_1(t) + N_i(t)}; \boldsymbol{\mu}\right), \end{aligned}$$

which concludes the case when  $i(t) = i$ .

**(3) When the other suboptimal arms are played** When  $i(t) \notin \{1, i\}$ ,  $N_1(t+1) = N_1(t)$  and  $N_i(t+1) = N_i(t) + 1$  holds. Therefore,  $(t+1)f_i(\mathbf{w}^{t+1}; \boldsymbol{\mu}) = tf_i(\mathbf{w}^t; \boldsymbol{\mu})$  holds, which concludes the case when  $i(t) \neq 1, i$ . □

## 5.7 Conclusion

In this chapter, we introduced BC-TE, a computationally efficient approach for solving the BAI problem in SPEF bandits. By combining a heuristic method with TS using the Jeffreys prior, BC-TE overcame the limitations of existing approaches that involve computationally expensive optimization problems, forced exploration steps, or hyperparameter tuning. Through theoretical analysis and experimental evaluation, we demonstrated

that TS with the Jeffreys prior can serve as a substitute for the forced exploration steps in BAI problems. Although BC-TE is not universally optimal in general, we showed its optimality for the two-armed bandits setting and provided a comparison with  $\beta$ -optimality. Here, BC-TE showed distinct benefits over  $\beta$ -optimal policies by eliminating the need for hyperparameter adjustments. Simulation results further validated the effectiveness of BC-TE, showing competitive sample complexity and improved computational efficiency compared to other optimal policies.

In summary, this chapter demonstrated the effectiveness of TS as an exploration mechanism in the BAI problem, providing a valuable contribution to solving the BAI problems. Although BC-TE may not achieve asymptotic optimality in all instances, it demonstrated unique advantages compared to other ( $\beta$ -)optimal policies across various criteria. Future work can focus on exploring asymptotically optimal policies with TE or investigating the application of this approach in more complex bandit settings, considering different problem constraints and assumptions.

## Chapter 6

### Conclusions and Future Work

In this chapter, we conclude this thesis by summarizing the contribution and introducing future research directions.

#### 6.1 Summary of the Thesis

While it is acknowledged that the choice of priors in Bayesian algorithms shares similarities with the selection of models for learning data in classification tasks, such as linear models or neural networks, the bandit literature has shown comparatively limited interest in exploring this aspect. In order to enhance the depth of this mathematically well-established field, it is essential to gain a comprehensive understanding of how the choice of priors impacts performance in bandit problems, akin to the consideration of priors in inference problems, rather than solely relying on a prior due to its simplicity or previous success in a specific statistical model.

It is important to note that the opinions expressed in the above discussion stem from the author’s own observations and beliefs. By presenting these opinions and exploring the role of priors in the context of Thompson sampling (TS), this thesis aims to foster further discussions and investigations in the area of sequential decision-making. Therefore, this thesis has been devoted to understanding how the choice of noninformative priors affects the behavior of TS in sequential decision-making problems with limited feedback.

In Chapter 3, we focused on the uniform model with unknown support, which can be seen as the simplest non-regular model. Through our investigation, we highlighted the importance of selecting appropriate noninformative priors by demonstrating the suboptimality of commonly used priors that are often considered natural choices. While the uniform prior on the location-scale parameterization achieves optimal performance in both the Gaussian and uniform models, we explicitly formalized the suboptimality of the uniform prior on alternative parameterizations, such as the location-rate parameterization where the rate parameter is defined as the inverse of the parameter. In particular, the uniform prior on the location-rate parameterization coincides with the Jeffreys prior in both models<sup>1</sup>, which exhibits suboptimal performance according to theoretical analysis and numerical experiments.

These results pointed out the necessity of an invariance property to handle general bandit models. In this context, we proposed a simple variant of TS, which we called TS with truncation (TS-T). By introducing an adaptive truncation procedure before sam-

---

<sup>1</sup>To be precise, the Jeffreys prior is not well-defined in the uniform model since the determinant of the Fisher information matrix of the uniform model is zero. However, for the purpose of consistency and to maintain a unified approach with the Gaussian model, we adopted the Jeffreys prior of the general location-scale family. It is worth noting that the reference prior, as demonstrated in Theorem 2.6 and Example 6, is well-defined and provides valuable insights in this context.

pling, we prevented the abrupt clustering of the posterior distribution in the early stage of learning. Remarkably, our results demonstrated that both the reference prior and the Jeffreys prior can achieve optimal performance when combined with TS-T, as validated by both theoretical analysis and numerical experiments.

Including the discussion in Chapter 3, the asymptotic optimality of TS has been considered only under the light-tailed distribution, and that for the heavy-tailed distribution remains unknown. To address this gap, we focused on the Pareto bandit model with unknown scale and shape parameters whose function is heavy-tailed in Chapter 4. The Pareto bandit model is of particular interest due to its relevance in analyzing web data [Mahanti et al., 2013] and social behavior patterns [Córdoba, 2008, Oancea et al., 2017], which have practical applications in online advertising.

Interestingly, we discovered the suboptimality problem exists not only for the reference prior and the Jeffreys prior but also for the uniform prior both on the scale-shape parameterization and on the rate-shape parameterization. These findings again emphasize the importance of selecting appropriate priors. However, we were able to establish the asymptotic optimality of the reference prior and the Jeffreys prior when combined with TS-T in the Pareto models. In short, we provided an alternative approach to achieving optimality:

Adding a truncation procedure to the parameter space of the posterior distribution instead of hunting for an optimal prior.

This approach has been shown to yield optimal performance in the Pareto models (non-regular heavy-tailed), the uniform models (non-regular compact), and the Gaussian models (regular light-tailed). It is important to note that the author does not claim that the proposed method can solve all problems optimally. However, it is expected to perform reasonably well when a fine-tuned truncation is applied, which would be much easier than finding optimal priors that yield computationally efficient posterior distributions.

The key contributions of Chapters 3 and 4 are summarized in Table 6.1. Throughout the thesis, we primarily focused on the optimality of the proposed approach using the reference priors due to their universal applicability in general statistical models and their favorable properties, as discussed in Section 2.3.

In Chapter 5, the best arm identification (BAI) problem with fixed confidence has been discussed, specifically addressing cases where the reward model belongs to the canonical single-parameter exponential family. Although the direct adaptation of TS is not suitable for this problem, we have developed a simple heuristic that combines TS with the Jeffreys prior as an exploration strategy, which naturally encourages exploration without the need for forced exploration techniques. From this combination, our method eliminates the need for solving computationally expensive optimization problems, forced exploration steps, and hyperparameter tuning. Therefore, we showed that TS could serve as a substitute for the traditional forced exploration steps, leading to improved efficiency and practicality. Furthermore, while our proposed method achieves near optimality, similar to the concept of  $\beta$ -optimality in Bayesian algorithms, we specifically showed its asymptotic optimality for two-armed bandit problems. This sets it apart from  $\beta$ -optimal policies, as our approach does not rely on the hyperparameter  $\beta$ , which offers distinct advantages. In short, we demonstrated the effectiveness of TS with the Jeffreys prior as an exploration mechanism in the BAI problems.

## 6.2 Future Directions

In this section, we discuss potential avenues for future research based on the findings and limitations presented in this thesis. While this thesis has contributed to improving

Table 6.1: Asymptotic optimality of TS with different noninformative priors for multiparameter models. R, C, and T denote whether the model satisfies the Fisher information regularity (✓) or not (✗), whether it is compact (✓) or non-compact (✗), and whether its function is light-tailed (L) or heavy-tailed (H). O-T and O-TT denote whether TS and TS-T can achieve the asymptotic lower regret bound in (2.1) for the corresponding model (✓) or not (✗), respectively. Notice that  $_H$  in O-T indicates that the results are derived by Honda and Takemura [2014].  $\pi_u$ ,  $\pi_j$ , and  $\pi_r$  denote the uniform prior for the specific parameterization in superscript, the Jeffreys prior, and the reference priors, respectively.

Model	R	C	T	Parameter $\theta$	Priors	O-T	O-TT
Gaussian Honda and Takemura [2014]	✓	✗	L	location (mean) and scale $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_u^{\mu, \sigma}$	✓ <sub>H</sub>	✓
					$\pi_j$	✗ <sub>H</sub>	✓
					$\pi_r$	✗ <sub>H</sub>	✓
Gaussian in Section 3.5				location (mean) and rate $(\mu, \sigma^{-1}) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_u^{\mu, \sigma^{-1}}$	✗	✓
					$\pi_j$	✗ <sub>H</sub>	✓
					$\pi_r$	✗ <sub>H</sub>	✓
Uniform in Chapter 3	✗	✓	L	location (mean) and scale $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_u^{\mu, \sigma}$	✓	✓
					$\pi_j$	✗	✓
					$\pi_r$	✗	✓
				location (mean) and rate $(\mu, \sigma^{-1}) \in \mathbb{R} \times \mathbb{R}_+$	$\pi_u^{\mu, \sigma^{-1}}$	✗	✓
					$\pi_j$	✗	✓
Pareto in Chapter 4	✗	✗	H	scale and shape $(\sigma, \alpha) \in \mathbb{R}_+ \times \mathbb{R}_{\geq 1}$	$\pi_r^{\sigma, \alpha}$	✗	✓
					$\pi_j$	✗	✓
					$\pi_r$	✗	✓

our understanding of the interplay between Bayesian algorithms and the selection of priors, there are still numerous theoretical and practical challenges that have not been fully addressed and thus require further investigation. These unexplored areas present exciting opportunities for future research directions.

### 6.2.1 Minimax optimality and asymptotic optimality

The main focus of this thesis has been on asymptotic optimality, which is for the expected regret. However, as we introduced in Section 2.1.1, the minimax optimality of the policy has been discussed in the bandit literature [Jin et al., 2021, Karbasi et al., 2021, Ménard and Garivier, 2017]. Recall that minimax optimality is valuable as it demonstrates the robustness of a policy across all possible bandit instances. However, it can sometimes provide overly loose bounds for specific instances where problem-dependent regret analysis offers more meaningful insights. We provide a simple illustration of the relationship between problem-dependent regret and problem-independent regret in Figure 6.1.

Therefore, it is very important to investigate whether TS can achieve both problem-dependent optimality (asymptotic optimality) and problem-independent optimality (minimax optimality) at the same time. It is worth noting that the minimax optimality of TS is currently only known for (sub-)Gaussian models [Jin et al., 2021] in the stochastic bandit problems. Therefore, we have the following future work.

**Problem 6.1.** *Prove whether TS(-T) with noninformative priors can achieve the asymptotic optimality and minimax optimality for the multiparameter models at the same time.*

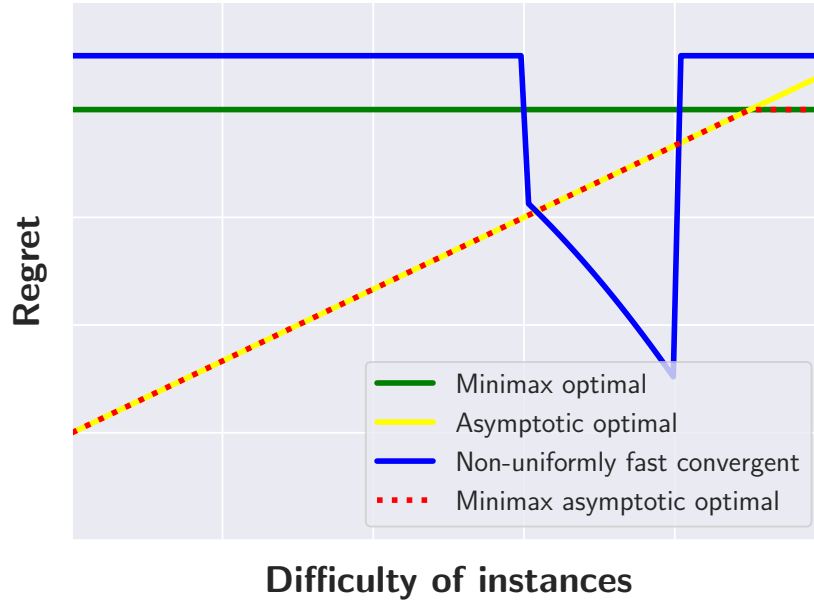


Figure 6.1: The relationship between the difficulty of an instance and the regret. Recall the definition of the uniformly fast convergent (UFC) policy in Definition 2.1, where Lai and Robbins [1985] and Burnetas and Katehakis [1996] showed that any UFC policies could not be below the “asymptotic optimal” line at any point. Note that it is not possible for any policy to be completely below the “minimax optimal” line. This illustration is inspired by Figure 16.1 in Lattimore and Szepesvári [2020].

### 6.2.2 Model misidentification

One of the biggest assumptions in the stochastic bandit problem is having prior knowledge about the reward model or family associated with each arm. This leads to a fundamental question:

What happens if the actual reward model deviates slightly or significantly from the assumed model?

Strictly speaking, it is important to acknowledge that our analysis for a specific model cannot provide any theoretical guarantee on the performance of TS when the actual model deviates, even in cases where TS for a Pareto model is applied to Pareto-like models that do not precisely follow Pareto distributions. Therefore, it becomes crucial to provide meaningful theoretical insights that remain applicable even under model misidentification or to design policies that are able to handle multiple models simultaneously.

#### Parametric models with many parameters

One straightforward approach to encompass various reward models is to increase the number of parameters, thereby enhancing the expressibility of more complex probability density functions. This increased expressibility enables the model to accommodate diverse reward models, providing greater adaptability and robustness in solving the stochastic bandit problem. In such cases, we expect the choice of the reference priors becomes more intuitive as they are originally designed to consider the case where the order of interest in parameters is different, and the bandit problem is the case where the primary focus is solely on the expected value of distributions. In this context, we can begin by considering the following problem, which is a simple but very practical setting.



**Problem 6.2.** *Prove whether TS(-T) with the reference prior can achieve the asymptotic optimality for the generalized Gaussian distributions, which is also known as the exponential power distributions.*

The optimal policy in this context can handle multiple distributions simultaneously since the model encompasses a wide range of commonly encountered distributions, including the uniform distribution, Gaussian distributions, and Laplace distributions. By considering this broader class of distributions, the policy can effectively adapt to various scenarios and provide robust performance across different distributional assumptions.

### **Best-of-both-worlds approach**

As we briefly introduced in Section 1.1.1, the assumptions on the reward model are eliminated in the adversarial bandit setting. While this setting represents an extreme scenario, achieving optimality in both the stochastic setting (without any assumptions on distributions) and the adversarial setting is an intriguing research direction. In this context, researchers have focused on developing best-of-both-worlds policies, which aim to perform well in both settings [Bubeck and Slivkins, 2012]. Recent research has made significant progress in this area, with notable contributions from Ito et al. [2022], Tsuchiya et al. [2023], and Zimmert and Seldin [2021].

However, the major challenge of the best-of-both-worlds policies is their computational cost since they often require solving optimization problems at each round. Notably, the Follow-The-Perturbed-Leader policy proposed by Honda et al. [2023] offers a promising exception to this issue. Based on these recent developments, one can consider the following directions.

**Problem 6.3.** *Develop computationally efficient best-of-both-worlds policies for general sequential decision-making problems.*

It is worth noting that the notion of optimality in the best-of-both-worlds setting for the stochastic reward setting is generally much looser compared to the optimality achieved for a specific reward model considered in this thesis. The difference in optimality between the best-of-both-worlds setting and the traditional stochastic setting highlights the trade-off between generality (robustness) and precision. While the best-of-both-worlds policies provide a more robust approach that performs reasonably well across different reward models, they may not achieve the same level of optimality as model-specific policies.

### **6.2.3 Theoretical derivation of the relationship between TS and priors**

Although this thesis aimed to provide intuition on how to choose a noninformative prior for certain models, it is important to note that a rigorous analysis of the relationship between TS and priors, applicable to a general model, still remains unaddressed. The focus of this thesis was primarily on specific reward models, where insights were provided based on the performance of TS with different priors. It is worth noting that the reference priors in the model considered in this thesis satisfy the probability matching property. Therefore, one can consider the following problem to establish a comprehensive and rigorous understanding of the relationship between TS and priors in a general model.

**Problem 6.4.** *What is the important factor of noninformative priors to understand the performance of TS: (i) the order of probability matching, (ii) maximizing missing information property, (iii) both, (iv) any others, or (v) no universal factors?*

## References

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.
- Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.
- Arvind Agarwal and Hal Daumé. A geometric view of conjugate priors. *Machine Learning*, 81:99–113, 2010.
- Priyank Agrawal, Jinglin Chen, and Nan Jiang. Improved worst-case regret bounds for randomized least-squares value iteration. In *The AAAI Conference on Artificial Intelligence*, volume 35, pages 6566–6573, 2021a.
- Rajeev Agrawal. Sample mean based index policies by  $o(\log n)$  regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.
- Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Annual Conference on Learning Theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.
- Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.
- Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.
- Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Annual Conference on Learning Theory*, pages 26–62. PMLR, 2021b.
- Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.
- FJ Anscombe. Sequential medical trials. *Journal of the American Statistical Association*, 58(302):365–383, 1963.
- Nicolás Aramayo, Mario Schiappacasse, and Marcel Goic. A multiarmed bandit approach for house ads recommendations. *Marketing Science*, 42(2):271–292, 2023.
- PMA Armitage. Sequential medical trials. *Sequential Medical Trials*, 1960.
- Barry C Arnold. Pareto and generalized Pareto distributions. In *Modeling income distributions and Lorenz curves*, pages 119–145. Springer, 2008.

- Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Annual Conference on Learning Theory*, volume 7, pages 1–122, 2009.
- Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 322–331. IEEE, 1995.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2):235–256, 2002.
- Vashist Avadhanula, Riccardo Colini Baldeschi, Stefano Leonardi, Karthik Abinav Sankararaman, and Okke Schrijvers. Stochastic bandits for multi-platform budget optimization in online advertising. In *International World Wide Web Conference*, pages 2805–2817, 2021.
- Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. In *Conference on Electronic Commerce*, pages 74–91, 2012.
- Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the Association for Computing Machinery*, 65(3):1–55, 2018.
- Antoine Barrier, Aurélien Garivier, and Tomáš Kocák. A non-asymptotic approach to best-arm identification for gaussian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 10078–10109. PMLR, 2022.
- Robert J Barro and Tao Jin. On the size distribution of macroeconomic disasters. *Econometrica*, 79(5):1567–1589, 2011.
- Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal Thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 716–726. PMLR, 2021.
- Gordon M Becker. Sequential decision making: Wald’s model and estimates of parameters. *Journal of Experimental Psychology*, 55(6):628, 1958.
- James O Berger and José M Bernardo. On the development of the reference prior method. *Bayesian Statistics*, 4(4):35–60, 1992.
- James O Berger, José M Bernardo, and Dongchu Sun. The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938, 2009.
- Jose M Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128, 1979.
- José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.
- David Blackwell and RV Ramamoorthi. A Bayes but not classically sufficient statistic. *The Annals of Statistics*, 10(3):1025–1026, 1982.
- Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *International Conference on Neural Information Processing*, pages 324–331. Springer, 2012.

- Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE, 2020.
- Brendan O Bradley and Murad S Taqqu. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pages 35–103. Elsevier, 2003.
- Björn Brodén, Mikael Hammar, Bengt J Nilsson, and Dimitris Paraschakis. Bandit algorithms for e-commerce recommender systems. In *The ACM Conference on Recommender Systems*, pages 349–349, 2017.
- Noel Bryson and Ayodele Mobolurin. An action learning evaluation procedure for multiple criteria decision making problems. *European Journal of Operational Research*, 96(2):379–386, 1997.
- Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and non-stochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for Thompson sampling. *Advances in Neural Information Processing Systems*, 26, 2013.
- Sébastien Bubeck and Mark Sellke. First-order bayesian regret analysis of thompson sampling. In *International Conference on Algorithmic Learning Theory*, pages 196–233. PMLR, 2020.
- Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Annual Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.
- Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, pages 23–37. Springer, 2009.
- Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.
- Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.
- Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.
- Giuseppe Burtini, Jason Loeppky, and Ramon Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*, 2015.
- T Tony Cai and Hongming Pu. Stochastic continuum-armed bandits with additive models: Minimax regrets and adaptive algorithm. *The Annals of Statistics*, 50(4):2179–2204, 2022.
- Junyu Cao and Wei Sun. Dynamic learning of sequential choice bandit problem under marketing fatigue. In *The AAAI Conference on Artificial Intelligence*, volume 33, pages 3264–3271, 2019.

- Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Annual Conference on Learning Theory*, pages 590–604. PMLR, 2016.
- Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Conference on Learning Theory*, pages 1295–1306. PMLR, 2018.
- Joel QL Chang and Vincent YF Tan. A unifying theory of Thompson sampling for continuous risk-averse bandits. In *The AAAI Conference on Artificial Intelligence*, volume 36, pages 6159–6166, 2022.
- Olivier Chapelle. Modeling delayed feedback in display advertising. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 1097–1105, 2014.
- Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, 24, 2011.
- Houshuang Chen, Shuai Li, and Chihao Zhang. Understanding bandits with graph feedback. *Advances in Neural Information Processing Systems*, 34:24659–24669, 2021.
- Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in Neural Information Processing Systems*, 27, 2014.
- Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159. PMLR, 2013.
- James Cheshire, Pierre Ménard, and Alexandra Carpentier. Problem dependent view on structured thresholding bandit problems. In *International Conference on Machine Learning*, pages 1846–1854. PMLR, 2021.
- Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch. On the frequency of severe terrorist events. *Journal of Conflict Resolution*, pages 58–87, 2007.
- Juan-Carlos Córdoba. On the distribution of city sizes. *Journal of Urban Economics*, 63, 2008.
- Wesley Cowan and Michael N Katehakis. An asymptotically optimal policy for uniform bandits of unknown support. *arXiv preprint arXiv:1505.01918*, 2015.
- Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *The Journal of Machine Learning Research*, 18(1):5638–5665, 2017.
- H Cramér. *Mathematical methods of statistics*. Princeton University Press, 1946.
- Will Dabney, Georg Ostrovski, and Andre Barreto. Temporally-extended  $\epsilon$ -greedy exploration. In *International Conference on Learning Representations*, 2021.
- Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Annual Conference on Learning Theory*. PMLR, 2008.
- Gauri Sankar Datta. On priors providing frequentist validity of Bayesian inference for multiple parametric functions. *Biometrika*, 83(2):287–298, 1996.

- Gauri Sankar Datta and Malay Ghosh. Some remarks on noninformative priors. *Journal of the American Statistical Association*, 90(432):1357–1363, 1995.
- Gauri Sankar Datta and Malay Ghosh. On the invariance of noninformative priors. *The Annals of Statistics*, 24(1):141–159, 1996.
- Gauri Sankar Datta and Rahul Mukerjee. *Probability Matching Priors: Higher Order Asymptotics: Higher Order Asymptotics*, volume 178. Springer Science & Business Media, 2004.
- Gauri Sankar Datta and Trevor J Sweeting. Probability matching priors. *Handbook of Statistics*, 25:91–114, 2005.
- R  my Degenne, Wouter M Koolen, and Pierre M  nard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019a.
- R  my Degenne, Thomas Nedelec, Cl  ment Calauz  nes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *International Conference on Artificial Intelligence and Statistics*, pages 1988–1996. PMLR, 2019b.
- R  my Degenne, Pierre M  nard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.
- Arnoud V Den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and Management Science*, 20(1): 1–18, 2015.
- Thomas J DiCiccio, Todd A Kuffner, and G Alastair Young. A simple analysis of the exact probability matching prior in the location-scale model. *The American Statistician*, 71(4):302–304, 2017.
- Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34: 1939–1951, 2021.
- Yihan Du, Yuko Kuroki, and Wei Chen. Combinatorial pure exploration with full-bandit or partial linear feedback. In *The AAAI Conference on Artificial Intelligence*, volume 35, pages 7262–7270, 2021.
- Abhimanyu Dubey and Alex Pentland. Thompson sampling on symmetric alpha-stable bandits. In *International Joint Conference on Artificial Intelligence*, 2019.
- Ward Edwards. The theory of decision making. *Psychological bulletin*, 51(4):380, 1954.
- Wedad Elmaghraby and Pinar Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science*, 49(10):1287–1309, 2003.
- Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, 2002.
- Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

- Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using Thompson sampling. *Operations Research*, 66(6):1586–1602, 2018.
- Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.
- Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.
- Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pages 3348–3356. PMLR, 2020.
- Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *IEEE Symposium on New Frontiers in Dynamic Spectrum*, pages 1–9. IEEE, 2010.
- Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260. PMLR, 2013.
- Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Annual Conference on Learning Theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.
- Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Annual Conference on Learning Theory*, 2016.
- Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016.
- Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *The Journal of Machine Learning Research*, 23(1):8049–8114, 2022.
- Malay Ghosh. Objective priors: An introduction for frequentists. *Statistical Science*, 26(2):187–202, 2011.
- John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.
- Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Annual Conference on Learning Theory*, pages 861–898. PMLR, 2015.
- Aditya Grover, Todor Markov, Peter Attia, Norman Jin, Nicolas Perkins, Bryan Cheong, Michael Chen, Zi Yang, Stephen Harris, and William Chueh. Best arm identification in multi-armed bandits with delayed feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 833–842. PMLR, 2018.
- Benjamin Han and Jared Gabor. Contextual bandits for advertising budget allocation. *Proceedings of the ADKDD*, 17, 2020.
- Botao Hao, Tor Lattimore, and Chao Qin. Contextual information-directed sampling. In *International Conference on Machine Learning*, pages 8446–8464. PMLR, 2022.

- Janis Hardwick and Quentin F Stout. Bandit strategies for ethical sequential allocation. *Computing Science and Statistics*, 23(6.1):421–424, 1991.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *The AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Roger M Hill. Applying Bayesian methodology with a uniform prior to the single period inventory model. *European Journal of Operational Research*, 98(3):555–562, 1997.
- Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Annual Conference on Learning Theory*, pages 67–79. PMLR, 2010.
- Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85:361–391, 2011.
- Junya Honda and Akimichi Takemura. Optimality of Thompson sampling for Gaussian bandits depends on priors. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2014.
- Junya Honda, Shinji Ito, and Taira Tsuchiya. Follow-the-perturbed-leader achieves best-of-both-worlds for bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 726–754. PMLR, 2023.
- Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical Bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.
- Ronald A Howard. Dynamic programming and Markov processes. 1960.
- Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Annual Conference on Learning Theory*, pages 1421–1422. PMLR, 2022.
- Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.
- Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck.  $\text{lil}'$  ucb: An optimal exploration algorithm for multi-armed bandits. In *Annual Conference on Learning Theory*, pages 423–439. PMLR, 2014.
- Kevin G Jamieson and Robert Nowak. Active ranking using pairwise comparisons. *Advances in Neural Information Processing Systems*, 24, 2011.
- Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.
- Harold Jeffreys. *Theory of probability and inference*. 3rd edition, 1961.
- Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mofs: Minimax optimal Thompson sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021.
- Marc Jourdan and Rémy Degenne. Non-asymptotic analysis of a ucb-based top two algorithm. *arXiv preprint arXiv:2210.05431*, 2022.



- Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. In *Advances in Neural Information Processing Systems*, 2022.
- Sham M Kakade, Adam Tauman Kalai, and Katrina Ligett. Playing games with approximation algorithms. In *The Annual ACM Symposium on Theory of Computing*, pages 546–555, 2007.
- Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, volume 87 of *PMLR*, pages 651–673. PMLR, 29–31 Oct 2018.
- Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, volume 12, pages 655–662, 2012.
- Amin Karbasi, Vahab Mirrokni, and Mohammad Shadravan. Parallelizing Thompson sampling. *Advances in Neural Information Processing Systems*, 34:10535–10548, 2021.
- Robert E Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.
- Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences*, 92(19):8584–8585, 1995.
- Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Annual Conference on Learning Theory*, pages 228–251, 2013.
- Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *The Journal of Machine Learning Research*, 22(1):11140–11183, 2021.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR, 2012a.
- Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012b.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of a/b testing. In *Annual Conference on Learning Theory*, pages 461–481. PMLR, 2014.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.
- LA Khalfin. Contribution to the decay theory of a quasi-stationary state. *Soviet Journal of Experimental and Theoretical Physics*, 6(6):1053–1063, 1958.
- Dal-Ho Kim, Sang-Gil Kang, and Woo-Dong Lee. Noninformative priors for Pareto distribution. *Journal of the Korean Data and Information Science Society*, 20(6):1213–1223, 2009.

- Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Annual Conference on Learning Theory*, pages 2328–2369. PMLR, 2020.
- Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Annual Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.
- Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Annual Conference on Learning Theory*, pages 1141–1154. PMLR, 2015.
- Junpei Komiyama, Taira Tsuchiya, and Junya Honda. Minimax optimal algorithms for fixed-budget best arm identification. *Advances in Neural Information Processing Systems*, 35:10393–10404, 2022.
- Wouter M Koolen. tidnabbil: Julia library for bandit experiments. <https://bitbucket.org/wmkoolen/tidnabbil/src/master/>, 2019.
- Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, 2013.
- Samuel Kotz and Norman L Johnson. *Breakthroughs in Statistics: Foundations and basic theory*. Springer Science & Business Media, 2012.
- Yuko Kuroki, Liyuan Xu, Atsushi Miyauchi, Junya Honda, and Masashi Sugiyama. Polynomial-time algorithms for multiple-arm identification with full-bandit feedback. *Neural Computation*, 2020.
- Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvári. Meta-Thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR, 2021.
- Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.
- Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.
- John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems*, 20(1):96–1, 2007.
- Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1820.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- Alessandro Lazaric and Emma Brunskill. Online stochastic optimization under correlated bandit feedback. In *International Conference on Machine Learning*, pages 1557–1565. PMLR, 2014.
- Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pages 6142–6151. PMLR, 2021.

- Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwa Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462, 2020.
- Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2nd edition, 2006.
- Mingming Li, Huafei Sun, and Linyu Peng. Fisher–Rao geometry and Jeffreys prior for Pareto distribution. *Communications in Statistics - Theory and Methods*, 2020.
- Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Annual Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.
- Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In *The AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.
- Kenneth S Lomax. Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49(268):847–852, 1954.
- Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Annual Conference on Learning Theory*, pages 1739–1776. PMLR, 2018.
- Aniket Mahanti, Niklas Carlsson, Anirban Mahanti, Martin Arlitt, and Carey Williamson. A tale of the tails: Power-laws in internet measurements. *IEEE Network*, 27(1):59–64, 2013.
- Henrick John Malik. Estimation of the parameters of the Pareto distribution. *Metrika*, 15(1):126–132, 1970.
- Benoit Mandelbrot. The Pareto-Levy law and the distribution of income. *International economic review*, 1(2):79–106, 1960.
- Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24, 2011.
- Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.
- Oden Maron and Andrew W Moore. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 11:193–225, 1997.
- Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 922–928. IEEE, 2015.
- Rita J Maurice. A minimax procedure for choosing between two populations using sequential sampling. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 19(2):255–261, 1957.

- Benedict C May, Nathan Korda, Anthony Lee, and David S Leslie. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13: 2069–2106, 2012.
- Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650. PMLR, 2016.
- Pierre Ménard. Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*, 2019.
- Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, 2017.
- Kanishka Misra, Eric M Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.
- Stefan Mittnik, Svetlozar T Rachev, and Marc S Paolella. Stable Paretian modeling in finance: Some empirical and theoretical aspects. *A practical guide to heavy tails*, pages 79–110, 1998.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- DS Moore. Maximum likelihood and sufficient statistics. *The American Mathematical Monthly*, 78(1):50–52, 1971.
- Jonas W Mueller, Vasilis Syrgkanis, and Matt Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. *Advances in Neural Information Processing Systems*, 32, 2019.
- Rahul Mukerjee and Malay Ghosh. Second-order probability matching priors. *Biometrika*, 84(4):970–975, 1997.
- Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, 1933.
- Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.
- Anna Nicolaou. Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):377–390, 1993.
- Makoto Nirei and Shuhei Aoki. Pareto distribution of income in neoclassical growth models. *Review of Economic Dynamics*, 2016.
- John P Nolan. *Univariate stable distributions*. Springer, 2020.

- Alessandro Nuara, Francesco Trovo, Nicola Gatti, and Marcello Restelli. A combinatorial-bandit algorithm for the online joint bid/budget optimization of pay-per-click advertising campaigns. In *The AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Bogdan Oancea, Tudorel Andrei, and Dan Pirjol. Income inequality in Romania: The exponential-Pareto distribution. *Physica A: Statistical Mechanics and its Applications*, 469, 2017.
- Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.
- Ian Osband, Benjamin Van Roy, Daniel J Russo, and Zheng Wen. Deep exploration via randomized value functions. *The Journal of Machine Learning Research*, 20(124): 1–62, 2019.
- Sandeep Pandey and Christopher Olston. Handling advertisements of unknown quality in search advertising. *Advances in Neural Information Processing Systems*, 19, 2006.
- HW Peers. On confidence points and bayesian probability points in the case of several parameters. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27 (1):9–16, 1965.
- Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.
- Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.
- Deniz Preil and Michael Krapp. Bandit-based inventory optimisation: Reinforcement learning in multi-echelon supply chains. *International Journal of Production Economics*, 252:108578, 2022.
- Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems*, 2017.
- Siddhartha Y Ramamohan, Arun Rajkumar, and Shivani Agarwal. Dueling bandits: Beyond Condorcet winners to general tournament solutions. *Advances in Neural Information Processing Systems*, 29, 2016.
- C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Resonance – Journal of Science Education*, 20:78–90, 1945.
- Charles Riou and Junya Honda. Bandit algorithms based on Thompson sampling for bounded reward distributions. In *International Conference on Algorithmic Learning Theory*, pages 777–826. PMLR, 2020.
- H Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55:527–535, 1952.
- Christian P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2nd edition, 2007.
- Christian P Robert, Nicolas Chopin, and Judith Rousseau. Rejoinder: Harold Jeffreys’ s theory of probability revisited. *Statistical Science*, 24(1):191–194, 2009.

- Aaron Roth, Aleksandrs Slivkins, Jonathan Ullman, and Zhiwei Steven Wu. Multidimensional dynamic pricing for welfare maximization. *ACM Transactions on Economics and Computation*, 8(1):1–35, 2020.
- Neil Rubens, Mehdi Elahi, Masashi Sugiyama, and Dain Kaplan. Active learning in recommender systems. *Recommender systems handbook*, pages 809–846, 2015.
- Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- Daniel Russo. Simple Bayesian algorithms for best arm identification. In *Annual Conference on Learning Theory*, 2016.
- Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.
- Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.
- Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.
- Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard. Indexed minimum empirical divergence for unimodal bandits. *Advances in Neural Information Processing Systems*, 34:7346–7356, 2021.
- Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. *Advances in Neural Information Processing Systems*, 25, 2012.
- Mark J Schervish. *Theory of statistics*. Springer Science & Business Media, 2012.
- Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.
- Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.
- Shahin Shahrampour, Mohammad Noshad, and Vahid Tarokh. On sequential elimination algorithms for best-arm identification in multi-armed bandits. *IEEE Transactions on Signal Processing*, 65(16):4281–4292, 2017.
- Xuedong Shang, Rianne Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, 2020.
- Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.
- Herbert A Simon. Theories of decision-making in economics and behavioral science. *The American Economic Review*, 49(3):253–283, 1959.
- M.K. Simon and D. Divsalar. Some new twists to problems involving the Gaussian probability integral. *IEEE Transactions on Communications*, 46(2):200–210, 1998. doi: 10.1109/26.659479.

- Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.
- Paul N Somerville. Some problems of optimum sampling. *Biometrika*, 41(3/4):420–429, 1954.
- Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.
- Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 197–207. University of California Press, 1956.
- Charles Stein. An example of wide discrepancy between fiducial and confidence intervals. *The Annals of Mathematical Statistics*, 30(4):877–880, 1959.
- Matthew Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. *Advances in Neural Information Processing Systems*, 21, 2008.
- Masashi Sugiyama. *Statistical reinforcement learning: modern machine learning approaches*. CRC Press, 2015.
- Fupeng Sun, Yueqi Cao, Shiqiang Zhang, and Huafei Sun. The Bayesian inference of Pareto models based on information geometry. *Entropy*, 2021.
- Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.
- Anne Randi Syversveen. Noninformative Bayesian priors. interpretation and problems with construction and applications. *Preprint statistics*, 3(3):1–11, 1998.
- Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3367–3375. PMLR, 2021.
- Mohammad Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, and Mikael Johansson. Stochastic online shortest path routing: The value of feedback. *IEEE Transactions on Automatic Control*, 63(4):915–930, 2017.
- Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. Personalized recommendation via parameter-free contextual bandits. In *Conference on Research and Development in Information Retrieval*, pages 323–332, 2015.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- Robert Tibshirani. Noninformative priors for one parameter of many. *Biometrika*, 76(3):604–608, 1989.
- Francesco Tricomi. Sulle funzioni ipergeometriche confluenti. *Annali di Matematica Pura ed Applicata*, 26(1):141–175, 1947.
- Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 1484–1515. PMLR, 2023.

- Taishi Uchiya, Atsuyoshi Nakamura, and Mineichi Kudo. Algorithms for adversarial bandit problems with multiple plays. In *International Conference on Algorithmic Learning Theory*, pages 375–389. Springer, 2010.
- MAJ Van Montfort and JV Witter. The generalized Pareto distribution applied to rainfall depths. *Hydrological Sciences Journal*, 31(2):151–162, 1986.
- Walter Vogel. A sequential design for the two armed bandit. *The Annals of Mathematical Statistics*, 31(2):430–443, 1960.
- A Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(4):117–186, 1945.
- David L. Wallace. Bounds on normal approximations to Student’s and the chi-square distributions. *The Annals of Mathematical Statistics*, 30(4):1121–1130, 1959.
- Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021a.
- Xikui Wang. Dynamic pricing with a poisson bandit model. *Sequential Analysis*, 26(4): 355–365, 2007.
- Yining Wang, Boxiao Chen, and David Simchi-Levi. Multimodal dynamic pricing. *Management Science*, 67(10):6136–6152, 2021b.
- George Neville Watson. *A treatise on the theory of Bessel functions*, volume 3. The University Press, 1922.
- Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Annual Conference on Learning Theory*, pages 1263–1291. PMLR, 2018.
- BL Welch and HW Peers. On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(2):318–329, 1963.
- Douglas J White. A survey of applications of markov decision processes. *Journal of the Operational Research Society*, 44(11):1073–1096, 1993.
- Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.
- Steven R Wilkinson, Cyrus F Bharucha, Martin C Fischer, Kirk W Madison, Patrick R Morrow, Qian Niu, Bala Sundaram, and Mark G Raizen. Experimental evidence for non-exponential decay in quantum tunnelling. *Nature*, 387(6633):575–577, 1997.
- Wei Xia and Roland Yap. Learning robust search strategies using a bandit-based approach. In *The AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 2018.
- Min Xu, Tao Qin, and Tie-Yan Liu. Estimation bias in multi-armed bandit algorithms for search advertising. *Advances in Neural Information Processing Systems*, 26, 2013.
- Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *International Joint Conference on Artificial Intelligence*, pages 2936–2942, 7 2020.



- Gyugeun Yoon and Joseph YJ Chow. Contextual bandit-based sequential transit route design under demand uncertainty. *Transportation Research Record*, 2674(5):613–625, 2020.
- Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.
- Houssam Zenati, Alberto Bietti, Eustache Diemert, Julien Mairal, Matthieu Martin, and Pierre Gaillard. Efficient kernelized ucb for contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 5689–5720. PMLR, 2022.
- Shunan Zhang and Michael D Lee. Optimal experimental design for a class of bandit problems. *Journal of Mathematical Psychology*, 54(6):499–508, 2010.
- Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Annual Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.
- Qian Zhou, XiaoFang Zhang, Jin Xu, and Bin Liang. Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*, pages 811–821. Springer, 2017.
- Qiuyu Zhu and Vincent Tan. Thompson sampling algorithms for mean-variance bandits. In *International Conference on Machine Learning*, pages 11599–11608. PMLR, 2020.
- Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.