OPTIMALITY, SCALABILITY, AND RENEGING IN BANDIT LEARNING

A Dissertation

by

XI LIU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	P. R. Kumar
Committee Members,	Xiaoning Qian
	Zhangyang Wang
	I-Hong Hou
Head of Department,	Miroslav Begovic

May 2020

Major Subject: Computer Engineering

ABSTRACT

Bandit learning has been widely applied to handle the exploration-exploitation dilemma in sequential decision problems. To solve the dilemma, a large number of bandit algorithms have been proposed. While many of these algorithms have been proved to be order-optimal with respect to regret, the difference between the best expected reward and that actually achieved, there remain two fundamental challenges.

First, the "efficiency" of the best-performing bandit algorithms is often unsatisfactory, where the efficiency is measured jointly with respect to the performance in maximizing rewards as well as the computational complexity. For instance, the Information Directed Sampling (IDS), variance-based IDS (VIDS), and Kullback-Leibler Upper Confidence Bounds (KL-UCB) have often been reported to achieve outstanding performance with respect to regret. Unfortunately, they suffer from high computational complexity even after approximation, and exhibit poor scalability of computational complexity as the number of arms increases. Second, most of the existing bandit algorithms assume that the sequential decision-making process will continue forever without an end. However, users may renege and stop playing. They also assume the underlying reward distribution is homoscedastic. Both these assumptions are often violated in real-world applications, where participants may disengage from future interactions if they do not have a rewarding experience, and at the same time, the variances of underlying distributions differs under different contexts.

To address the aforementioned challenges, we propose a family of novel bandit algorithms. To address the efficiency issue, we propose Biased Maximum Likelihood Estimation (BMLE) - a family of novel bandit algorithms that generally apply to both parametric and non-parametric reward distributions, often have a closed-form solution and low computation complexity, have a quantifiable regret bound, and demonstrate satisfactory empirical performance. To enable bandit algorithms handle the reneging risk and reward heteroscedasticity, we propose a Heteroscedastic Reneging Upper Confidence Bound policy (HR-UCB) - a novel UCB-type algorithm that achieves outstanding and quantifiable performance in the presence of reneging risk and heteroscedasticity.

DEDICATION

To my mom Xiuyun Zhang and dad Kaiyuan Liu.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor P. R. Kumar, for his guidance and support throughout my PhD study. He opened up my mind in determining research problems and formulating meaningful models for them. I have been continuously inspired by his creative thinking, profound advice, and insightful ideas about both research and life.

I also want to acknowledge Professors Xiaoning Qian, Zhangyang Wang, and I-Hong Hou for serving on my dissertation committee. Their inspiring suggestions have significantly improved the quality of this dissertation.

I had a great fortune to be hosted by Dr. Rui Chen and Prof. Yong Ge in Samsung Research. Special thanks to my colleague and roommate Ping-Chun Hsieh, for taking an adventure in both research and foods with me. Both of us enjoyed ourselves and gained considerable knowledge as well as weight from the adventure.

Special thanks to my girlfriend Xinru Ma, who has been taking care of me during the period of my graduate study. Without her tolerance and encouragement, I would not be able to achieve those milestones over the years.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor P. R. Kumar (academic advisor), Professor Xiaoning Qian and Professor I-Hong Hou, of the Department of Electrical and Computer Engineering, and Professor Zhangyang Wang of the Department of Computer Science and Engineering.

The proofs in Sections 2 and 3 were conducted in collaboration with Ping-Chun Hsieh of Electrical and Computer Engineering. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by NSF under Contract Nos. CNS-1646449, CPS-1239116, CCF-1619085 and Science & Technology Center Grant CCF0939370, and Texas A&M University under the Presidents Excellence Funds X Grants Program.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Fundamental Challenges in Bandit Learning

Bandit learning has been widely applied to solve sequential decision problems in many applications such as web advertising, recommender systems, information retrieval, clinical trials, etc. The classical difficulty addressed in bandit learning is the exploration-exploitation dilemma, which requires the learning algorithm to balance information gathering and best use of available information to achieve optimal performance. Many bandit algorithms have been proposed to overcome this difficulty in the existing literature. These algorithms can be categorized into two main groups: frequentist approaches (e.g., UCB [1], UCB-Tuned [1], MOSS [2, 3], KL-UCB [4, 5, 6]) and Bayesian approaches (e.g., Bayes-UCB [7], Thompson sampling [8, 9, 10, 11], Knowledge Gradient [12, 13] Information Directed Sampling [14, 15]). In the frequentist settings, an upper confidence bound is derived from concentration inequalities or constructed with the help of other information measures, such as the Kullback–Leibler divergence. The Bayesian approaches, on the other hand, assume that the unknown parameters are drawn from an underlying prior distribution, and make the decisions by following a continually updated posterior distribution.

While many algorithms from both groups have been proved to be order-optimal, there are two fundamental challenges that are inadequately addressed in the existing literature. First, the "efficiency" of the best-performing algorithms is often unsatisfactory. Here efficiency is measured in terms of the performance in maximizing reward accumulation with respect to computational complexity. For instance, it is well known that among frequentist approaches, although UCB, UCB-Tuned, and MOSS have a closed-form solution and low computational complexity, their empirical performance is often worse than KL-UCB, which has no closed-form solution and has higher computational complexity. Among Bayesian approaches, Information Directed Sampling (IDS) and its variant variance-based IDS (VIDS) have often demonstrated state-of-the-art performance compared to Thompson sampling, Bayes-UCB, and Knowledge Gradient. Unfortunately, IDS has no

closed-form solution and suffers from high computational complexity even after approximation and poor scalability with a large number of arms. This limitation restricts the application scope of bandit learning in large-scale machine learning problems, where efficiency and scalability are major concerns.

Second, most of these algorithms are designed to handle sequential decision problems that continue indefinitely without an end. In addition, they also assume that the unknown environment has homoscedastic reward distributions, i.e., the variance is the same across different reward distributions to be learned. However, these assumptions are often violated in real-world applications such as clinical trials, portfolio selection, and cloud computing. In these applications, participants may disengage from future interactions if they receive insufficient rewards, and at the same time, the reward distributions have been observed to be heteroscedastic [16, 17, 18, 19, 20]. These violations may render nominally optimal algorithms sub-optimal and face difficulty in sustaining their outstanding performance in these applications.

1.2 Outline of the Dissertation

To address the aforementioned challenges in efficiency, we propose a family of novel bandit algorithms in Section 2. To address the efficiency issue, we propose Biased Maximum Likelihood Estimation (BMLE) - a family of novel bandit algorithms that can be generally applied to reward distributions from both parametric families (e.g., exponential family) as well as non-parametric families (e.g., sub-Gaussian and sub-exponential). Compared to existing bandit algorithms, BMLE has several salient features. First, it has a closed-form solution and low computational complexity, demonstrating promising scalability with a large number of arms. Second, the regret bound of BMLE is quantifiable and is order-optimal under mild assumptions. Finally, it often demonstrates outstanding empirical performance along with a major computational advantage in comparison to many other state-of-the-art methods.

To enable bandit algorithms to handle learning tasks with reneging risk and heteroscedastic rewards, in Section 3, we propose Heteroscedastic Reneging Upper Confidence Bound algorithm (HR-UCB) - a novel UCB-type bandit algorithm that is able to work in the presence of reneg-

ing phenomena and heteroscedastic reward distributions. We prove a regret bound for HR-UCB and evaluate its performance in comprehensive experiments. We find that the performance of existing methods such as LinUCB, Contextual Markov Decision Process (CMDP), and Episodic Reinforcement Learning (ERL) is unsatisfactory, while HR-UCB demonstrates excellent empirical performance. We provide concluding remarks in Section 4.

For better readability, the detailed proofs for Section 2 and Section 3 are provided in Appendix A and Appendix B respectively. In the main bodies of the two sections, intuition and sketches of the proofs are given.

2. OPTIMALITY AND SCALABILITY OF STOCHASTIC MULTI-ARMED BANDITS THROUGH BIASED MAXIMUM LIKELIHOOD ESTIMATION[1]

2.1 Overview

In this section, we introduce BMLE – a family of novel learning algorithms for stochastic multi-armed bandit problems. Algorithm design for the stochastic multi-armed bandit problem has been studied extensively in the literature. Most prior work can be categorized into two main groups, namely frequentist approaches and Bayesian approaches. Frequentist approaches consider the unknown reward parameters as fixed but unknown. An optimistic estimate (empirical mean plus confidence bound) of the unknown parameters is relied upon to guide the sequential decisions. The family of Upper Confidence Bound (UCB) algorithms is among the most popular in this group, given its simplicity in implementation and good theoretical guarantees. In this family, UCB, UCB-Tuned, and MOSS directly construct their upper confidence bound from concentration inequalities and have a closed-form solution [22, 1, 23, 2, 3]. In comparison, KL-UCB derives the bound with the help of other information measures, such as the Kullback-Leibler divergence, and has no closed-form solution [4, 5, 7, 6]. On the other hand, Bayesian approaches consider the unknown reward parameters to have been drawn from an underlying prior distribution. As rewards are accumulated, algorithms in this group continually update and base decisions on the posterior distribution of the unknown parameters. In this family, Thompson sampling, Knowledge Gradient (KG), KG* and Bayes-UCB directly apply the statistics of the updated posterior distribution and thus have a closed-form solution [24, 8, 9, 11, 10, 12, 13, 7]. In contrast, information-directed sampling (IDS) blends the concept of information gain by looking at the ratio between the square of expected immediate regret and the expected reduction in the entropy of the target, and has no closed-form solution [14, 15].

While many algorithms from both groups have been proved to be order-optimal, one funda-

[1]Part of this section is reprinted from my preprint "Bandit Learning Through Biased Maximum Likelihood Estimation" by Xi Liu, Ping-Chun Hsieh, Anirban Bhattacharya, and P. R. Kumar [21] that is publicly available at https://arxiv.org/abs/1907.01287

mental limitation is that their "efficiency" is often unsatisfactory. Here efficiency refers to their performance in maximizing reward accumulation with respect to computational complexity in making decisions. For instance, it is well known that among the frequentist approaches, although UCB, UCB-Tuned, and MOSS have a closed-form solution and low computational complexity, their empirical performance is often worse than KL-UCB, which has no closed-form solution but has higher computational complexity [7, 6]. Not surprisingly, among Bayesian approaches, the Information Directed Sampling (IDS) has demonstrated state-of-the-art best performance compared to Thompson sampling, Bayes-UCB, and Knowledge Gradient. This statement is even true when comparing IDS with frequentist approaches [14, 15]. Unfortunately, IDS and its variant V-IDS have no closed-form solution and suffer from high computational overhead due to the excessive sampling required for estimating the integrals involved. This limitation restricts the applicability of state-of-the-art bandit learning algorithms in large-scale machine learning problems, where efficiency is a significant concern, e.g., when the number of arms is in the billions, KL-UCB and IDS are unscalable. Another issue with respect to Bayesian approaches is that their performance has been reported to be sensitive to the choice of prior [11, 25].

To attack the aforementioned challenges, we revisit the bandit learning problem from the frequentist perspective and propose Biased Maximum Likelihood Estimation (BMLE) - a family of novel bandit algorithms that can be generally applied to reward distributions from both parametric families (e.g., exponential family) as well as non-parametric families (e.g., distributions with bounded support). Compared to existing bandit algorithms, BMLE has several salient features. First, BMLE does not rely on a prior and hence completely obviates the potential issues arising from an inappropriate choice of prior. Second, BMLE has a closed-form solution and low computational complexity. Third, the regret bound of BMLE is quantifiable, and it is order-optimal under mild assumptions. Finally, it often demonstrates empirical performance comparable to the best, but at the same time, retain computational efficiency. As a result, in large-scale machine learning problems, BMLE may be preferred in comparison to other baseline schemes. The intuition that BMLE outperforms its counterparts in the frequentist framework is that most of those base-

lines rely on the upper confidence bound to construct the index. However, the upper confidence bound only uses moment assumptions on the true distribution and hence does not fully exploit all underlying information. In contrast, the proposed BMLE algorithm addresses the exploration and exploitation trade-off by directly operating with the likelihood function to navigate the exploration. This feature allows it to makes better use of the information on the parametric distributions. As such, BMLE can provide simple new indices for bandits with well-known distributions. These indices are quite different from, for example, UCB-based indices. Table 2.1 shows a comparison of the indices produced by BMLE and other UCB-based policies. The fact that such qualitatively different indices provide excellent performance may itself be of intrinsic interest.

Algorithm	Index
BMLE	(Bernoulli) $N_i(t)\big(H(p_i(t)) - H(\tilde{p}_i(t))\big)$
	(Gaussian) $p_i(t) + \alpha(t)/(2N_i(t))$
	(Exponential) $N_i(t)\log\left(\frac{N_i(t)p_i(t)}{N_i(t)p_i(t)+\alpha(t)}\right)$
UCB	$p_i(t) + \sqrt{2\log t/N_i(t)}$
UCB-Tuned	$p_i(t) + \sqrt{\min\{\frac{1}{4}, \overline{V}_t(i)\}\log(t)/N_i(t)\}}$
MOSS	$p_i(t) + \sqrt{\max(\log(\frac{T}{N_i(t)\cdot N}), 0)/N_i(t)}$

Table 2.1: Comparison of indices produced by BMLE with other approaches. Below $H(p)$ is the binary entropy, $\overline{V}_t(i)$ is the upper bound on the variance, and the other quantities are defined in Sections 2.3 and 2.4.2.

The main contributions of this section are as follows:

- We present a new family of bandit algorithms from the perspective of biased maximum likelihood estimation.
- We substantiate the BMLE algorithm by considering the general exponential family reward distributions. By designing proper bias terms for the likelihood function, we derive simple closed-form new indices for different bandit problems.
- For both Gaussian distributions and other exponential families that satisfy some mild conditions, we provide the first logarithmic regret bound for the BMLE algorithm and thereby

characterize the interplay between the bias term and the regret. The same regret bounds also extend to non-parametric reward distributions.

- We conduct extensive numerical simulations and show that the BMLE algorithm can achieve or better state-of-the-art regret performance. Through extensive comparative numerical simulation of several competitive algorithms, we also establish the efficiency of BMLE in terms of computational time per pull and scalability in terms of the number of arms.

2.2 Related Work

The algorithm design for the stochastic multi-armed bandit problem has been studied extensively in the existing literature. Most of the prior work can be categorized into two main groups, namely frequentist approaches and Bayesian approaches. In the frequentist settings, the family of UCB algorithms, including UCB [1], UCB-Tuned (UCBT) [1], and MOSS [2, 3], are among the most popular ones given their simplicity in implementation and good theoretical guarantees. An upper confidence bound can be directly derived from concentration inequalities or constructed with the help of other information measures, such as the Kullback–Leibler divergence used by the KL-UCB algorithm [4, 5, 7, 6]. The concept of upper confidence bound has later been extended to various types of models, such as contextual linear bandits [26, 27, 28], Gaussian process bandit optimization [29], and model-based reinforcement learning [30]. The above list is by no means exhaustive but is mainly meant to illustrate the wide applicability of the UCB approach in different settings. While being a simple and generic index-type algorithm, UCB-based methods sometimes suffer from much higher regret than their counterparts [14, 8]. This mainly results from the fact that the upper confidence bound itself only uses moment assumptions on the true distribution and hence does not fully exploit the underlying information structure. Different from the UCB solutions, the proposed BMLE algorithm addresses the exploration and exploitation trade-off by directly operating with the likelihood function to navigate the exploration, and therefore it makes better use of the information of the parametric distributions.

On the other hand, the Bayesian approach studies the setting where the unknown reward pa-

rameters are drawn from an underlying prior distribution. As one of the most popular Bayesian bandit algorithms, Thompson sampling (TS) [24, 8, 9, 11, 10] follows the principle of probability matching by continuously updating the posterior distribution based on a prior. In addition to strong theoretical guarantees, [9, 10], TS has been reported to achieve superior empirical performance to its counterparts [8, 24]. While being a powerful bandit algorithm, TS can be sensitive to the choice of the prior [11, 25]. Another popular Bayesian algorithm is Bayes-UCB [7], which combines the Bayesian interpretation of bandit problems and the simple closed-form expression of UCB-type algorithms. In contrast, BMLE does not rely on a prior and hence completely obviates the potential issues arising from an inappropriate prior choice.

Another line addresses the exploration and exploitation dilemma through information-related measures. The Knowledge Gradient (KG) approach [12] and its variant KG* [13], KGMin, and KGMN [31, 12, 13] proceed by making a greedy one-step look-ahead measurement for exploration, as suggested by their name. While KG has been shown empirically to perform well for Gaussian process optimization [13, 32], its performance is not readily quantifiable, and it does not always converge to optimality [14]. Another promising solution is the Information Directed Sampling (IDS) and its variant - VIDS [14, 15] proposed by Russo and Van Roy [14, 15]. Different from the KG algorithm, IDS blends in the concept of information gain by looking at the ratio between the square of expected immediate regret and the expected reduction in the entropy of the target. Moreover, it has been reported in [14, 15] that IDS achieves state-of-the-art results in various bandit models. However, IDS and its variants can suffer from high computational time per decision (i.e., pull) due to the excessive sampling required for estimating high dimensional integrals. Compared to these competitive solutions, the proposed BMLE method can achieve comparable performance both theoretically and empirically, but at the same time retains computational efficiency.

Our work also connects to adaptive control of unknown MDPs. The stochastic N-armed bandit problem, in general, can be viewed as an unknown MDP problem. This has historically been a challenging problem [33] since an action taken on a dynamic system serves the "dual" purposes [34, 35] of controlling the system to reduce the immediate cost incurred and also simultaneously explor-

ing system behavior by exciting it. A straightforward solution of the problem is estimating the unknown parameter, and then taking an action that would be optimal for the estimate, which is often referred to as the "certainty equivalence" approach. However, this approach suffers from the "closed-loop identifiability" problem [36]: the system is ever-evolving in a closed-loop with the adaptive control law, and as the control law converges to limiting control law, it ceases to learn about other possibly better control laws [36, 37, 38, 39].

Specifically, consider a Markov Decision Process with state space X, action space U, with controlled transition probabilities $p(i, j, u; \boldsymbol{\theta})$ denoting the probability of transition to a next state $j \in X$ when the current state is $i \in X$ and action $u \in U$ is applied, indexed by a parameter $\boldsymbol{\theta}$ in a set Θ. The true parameter is $\boldsymbol{\theta}^0 \in \Theta$, but is unknown. A reward $r(i, j, u)$ is accrued when the system transitions from i to j under u. The goal is to maximize the long-term average reward $\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c(x(t), x(t+1), u(t))$, where $x(t)$ and $u(t)$ are the state and action taken at time t. Let $\phi_{\boldsymbol{\theta}} : X \to U$, be a stationary control law such that choosing $u(t) = \phi_{\boldsymbol{\theta}}(x(t))$ is optimal if the true parameter is $\boldsymbol{\theta}$; such an optimal control law exists under various conditions [40]. Since the true parameter $\boldsymbol{\theta}^0$ is unknown, one can employ a certainty-equivalent strategy of making a maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}(t) \in \Theta$ that maximizes the likelihood $\prod_{s=0}^{t-1} p(x(s), x(s+1), u(s), \boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$, and then applying the action $u(t) = \phi_{\hat{\boldsymbol{\theta}}(t)}(x(t))$. Then, the parameter estimates $\hat{\boldsymbol{\theta}}(t)$ converge to a $\boldsymbol{\theta}^*$ such that

$$p(i, j, \phi_{\boldsymbol{\theta}^*}(i), \boldsymbol{\theta}^*) = p(i, j, \phi_{\boldsymbol{\theta}^*}(i), \boldsymbol{\theta}^0) \text{ for all } i, j. \tag{2.1}$$

However $\phi_{\boldsymbol{\theta}^*}$ need not be optimal for $\boldsymbol{\theta}^0$.

A solution to this fundamental problem was proposed in [41]. Let $J(\phi, \boldsymbol{\theta})$ denote the long-term average reward accrued by the stationary control law ϕ, and $J_{opt}(\boldsymbol{\theta}) := \max_\phi J(\phi, \boldsymbol{\theta})$ the optimal long-term reward, when the parameter is $\boldsymbol{\theta}$. Then the closed-loop identification (2.1) implies that $J(\phi_{\boldsymbol{\theta}^*}, \boldsymbol{\theta}^*) = J(\phi_{\boldsymbol{\theta}^*}, \boldsymbol{\theta}^0)$. However, since $J(\phi_{\boldsymbol{\theta}^*}, \boldsymbol{\theta}^*) = J_{opt}(\boldsymbol{\theta}^*)$, while $J(\phi_{\boldsymbol{\theta}^*}, \boldsymbol{\theta}^0) \leq J_{opt}(\boldsymbol{\theta}^0)$, it follows that $J_{opt}(\boldsymbol{\theta}^*) \leq J_{opt}(\boldsymbol{\theta}^0)$. Therefore it was suggested in [41] to introduce a delicate

bias into the maximum likelihood estimate to prefer parameters with higher optimal rewards. The resulting "biased maximum likelihood estimation" (BMLE) is:

$$\hat{\boldsymbol{\theta}}^{\text{BMLE}}(t) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, J(\boldsymbol{\theta})^{\alpha(t)} \prod_{s=0}^{t-1} p(x(s), x(s+1), u(s), \boldsymbol{\theta}),$$

where $\alpha(t) : [1, \infty) \to \mathbb{R}_+$ is a function that satisfies $\lim_{t\to\infty} \alpha(t) = \infty$ and $\lim_{t\to\infty} \alpha(t)/t = 0$. The control action chosen is $u(t) = \phi_{\hat{\boldsymbol{\theta}}^{\text{BMLE}}(t)}(t)(x(t))$. The cost-bias term $J(\boldsymbol{\theta})^{\alpha(t)}$ has two salient features: (i) $J(\boldsymbol{\theta})^{\alpha(t)}$ achieves active exploration by favoring models with higher reward, and (ii) the effect of the bias term gradually diminishes as $\alpha(t)$ grows indefinitely with time. This method was shown to yield the optimal long-term average reward in a variety of settings [42, 43, 44, 45, 46, 47, 48, 49, 50]. Long-term average optimality studied in the BMLE work is a gross measure implying only that *regret* is $o(t)$. However, in bandit learning [22], attention has been focused on showing a much finer $O(\log(t))$ optimality of regret. No existing study has addressed whether and how the cost-bias idea still works in bandit learning, where guarantees on finite-time performance are indispensable, and where, therefore, a finer measure of optimality is of interest. As such, our goal in this section is to tailor the BMLE to the stochastic multi-armed bandit problem and perform finite-time analysis in terms of *regret*.

2.3 Problem Formulation

We consider the stochastic N-armed bandit problem, where each arm i is characterized by its reward distribution $\mathcal{D}_i$ with mean θ_i. Without loss of generality, we assume that $\theta_1 > \theta_2 > \cdots > \theta_N \geq 0$, and hence arm 1 is the optimal arm. For each arm i, we define $\Delta_i := \theta_1 - \theta_i$ to be the negative of the gap between its mean reward and that of the optimal arm. For ease of notation, we also use Δ to denote the minimum gap of Δ_2. We use $\boldsymbol{\theta}$ to denote the vector $(\theta_1, \cdots, \theta_N)$. At each time $t = 1, \cdots, T$, the decision maker chooses an arm $\pi_t \in \{1, \cdots, N\}$ and obtains a corresponding reward X_t, which is independently drawn from the distribution $\mathcal{D}_{\pi_t}$. Let $N_i(t)$ and $S_i(t)$ be the total number of trials of arm i and the total reward collected from pulling arm i up to time t, respectively. We also use $\mathcal{H}_t = (\pi_1, X_1, \pi_2, X_2, \cdots, \pi_t, X_t)$ to denote

the history of all the choices of the decision maker and the reward observations up to time t. We let $L(\mathcal{H}_t; \{\mathcal{D}_i\})$ denote the likelihood of the history $\mathcal{H}_t$ under the reward distributions $\{\mathcal{D}_i\}$. Based on the multi-armed bandit convention, our objective is to minimize the *pseudo regret* defined as $\text{Regret}(T) := T\theta_1 - \mathbb{E}[\sum_{t=1}^{T} X_t]$, where the expectation is taken with respect to the randomness of the rewards and the employed policy. The employed policy should not depend on T and should perform well for all T.

2.4 The BMLE Algorithm

In this section, we formally introduce the general procedure of BMLE and then substantiate it by considering a collection of commonly-studied parametric reward distributions.

2.4.1 The Generic BMLE Procedure

The main components of the BMLE algorithm are:

- Design a bias term that favors the models with larger achievable optimal long-term average reward.
- At each time t, derive the biased maximum likelihood estimator $\hat{\boldsymbol{\theta}}_t^{\text{BMLE}} = (\hat{\theta}_{t,i}^{\text{BMLE}})$ as detailed in the subsequent subsections, and then select an arm as

$$\pi_t^{\text{BMLE}} = \underset{i\in\{1,\cdots,N\}}{\text{argmax}}\ \hat{\theta}_{t,i}^{\text{BMLE}}. \tag{2.2}$$

  (We assume throughout that some arbitrary order on the argument of "argmax" is used to break ties).

2.4.2 BMLE Index For Exponential Family Distributions

In this section, we discuss the BMLE algorithm for the exponential family reward distributions. To begin with, the probability density or mass function of an exponential family distribution in natural form can be expressed as

$$p(x;\eta) = A(x)\exp\big(\eta x - F(\eta)\big),\ \eta \in \mathcal{N} \tag{2.3}$$

where η is the canonical parameter, $\mathcal{N}$ is the parameter space, $A(\cdot)$ is a real-valued function, and $F(\cdot)$ is a real-valued twice-differentiable function. For example, a Gaussian distribution with mean μ_i and known variance σ_i^2 can be represented in the form of (2.3) by letting $\eta = \mu_i/\sigma_i^2$, $F(\eta) = \sigma_i^2\eta^2/2$, $A(x) = (2\pi\sigma_i^2)^{-\frac{1}{2}}\exp(-x^2/2\sigma_i^2)$. By calculating the moment generating function for $p(x;\eta)$, we further know that $\mathbb{E}[Y] = \dot{F}(\eta)$ ("dot" denoting derivative) and $\text{Var}[Y] = \ddot{F}(\eta)$, for any random variable Y with a density function as (2.3) This also suggests that $F(\eta)$ is strictly convex and the mean function $\dot{F}(\eta)$ is strictly increasing. Therefore, there is a one-to-one mapping between the canonical parameter and the mean parameter. We use $\dot{F}^{-1}(\cdot)$ to denote the inverse function of $\dot{F}(\cdot)$. Moreover, we use $\text{KL}(\eta' \,||\, \eta'')$ to denote the Kullback-Leibler (KL) divergence of any two distributions in an exponential family with canonical parameters η' and η''. In an exponential family, the KL divergence can be further expressed as

$$\text{KL}(\eta' \,||\, \eta'') = F(\eta'') - [F(\eta') + \dot{F}(\eta')(\eta'' - \eta')]. \tag{2.4}$$

Given that there is a one-to-one mapping between the canonical parameter and the mean parameter in an exponential family, we further define $D(\theta', \theta'') : \Theta \times \Theta \to \mathbb{R}_+$ as

$$D(\theta', \theta'') := \text{KL}(\dot{F}^{-1}(\theta') \,||\, \dot{F}^{-1}(\theta'')). \tag{2.5}$$

Next, we turn to the derivation of the proposed BMLE index. Consider the case where the reward distribution of each arm i has the density function $p(x;\eta_i)$ with mean $\theta_i = \dot{F}(\eta_i)$, and $F(\cdot)$ and $A(\cdot)$ are identical across all the arms. We use $\boldsymbol{\eta}$ to denote the vector $(\eta_1, \cdots, \eta_N)$. Recall that π_t denotes the index of the arm chosen by the employed policy at time t. Based on (2.3), we know that at each time t, the likelihood of $\mathcal{H}_t$ under the parameters $\boldsymbol{\eta}$ of the distribution is

$$L(\mathcal{H}_t; \boldsymbol{\eta}) = \prod_{s=1}^{t} A(X_s)\exp\left(\eta_{\pi_s}X_s - F(\eta_{\pi_s})\right). \tag{2.6}$$

Next, we propose to construct the multiplicative bias term as $\max_{i\in\{1,\cdots,N\}}\exp(g(\dot{F}(\eta_i))\,\alpha(t))$,

where $g(\cdot)$ is a strictly increasing user-defined real-valued function. We specifically choose $g(\cdot)$ to be the inverse function of $\dot{F}(\cdot)$ and hence $g(\dot{F}(\eta)) = \eta$. Then, the BMLE index for exponential family distributions can be derived as

$$\hat{\boldsymbol{\eta}}_t^{\text{BMLE}} := \underset{\eta_i \in \mathcal{N}, \forall i}{\operatorname{argmax}} \big\{ L(\mathcal{H}_t; \boldsymbol{\eta}) \max_{1 \le i \le N} \exp(\eta_i \alpha(t)) \big\}. \tag{2.7}$$

The relationship between (2.7) and BMLE is as follows: Each $\boldsymbol{\eta}$ represents an instance of the bandit model with mean reward equal to $\dot{F}(\eta_i)$ for each arm i. Under each instance $\boldsymbol{\eta}$, the optimal long-term average reward is simply $\max_{1 \le i \le N} \dot{F}(\eta_i)$. Since $\dot{F}(\cdot)$ is a strictly increasing function and $\theta_i = \dot{F}(\eta_i)$, we may use η_i as a proxy of the mean reward θ_i in designing the bias term. Therefore, with a positive function $\alpha(t)$, $(\max_{1 \le i \le N} \exp(\eta_i \alpha(t))$ is indeed a bias term in favor of the models with larger achievable optimal long-term average reward.

Next, we derive a simple closed-form expression for π_t^{BMLE}. Based on (2.2) and the maximization problem of (2.7), we know

$$\pi_t^{\text{BMLE}} \tag{2.8}$$

$$= \underset{i \in \{1, \cdots, N\}}{\operatorname{argmax}} \underset{\eta_i \in \mathcal{N}, \forall i}{\operatorname{argmax}} \big\{ \max_{1 \le i \le N} L(\mathcal{H}_t; \boldsymbol{\eta}) \exp(\eta_i \alpha(t)) \big\} \tag{2.9}$$

$$= \underset{i \in \{1, \cdots, N\}}{\operatorname{argmax}} \{ \max_{\eta_i \in \mathcal{N}, \forall i} L(\mathcal{H}_t; \boldsymbol{\eta}) \exp(\eta_i \alpha(t)) \}, \tag{2.10}$$

where (2.10) is obtained by exchanging the order of the two inner maximizations in (2.9). By solving the inner maximization problem in (2.10), we show that BMLE enjoys a simple closed-form expression for the exponential families. Define

$$\begin{aligned} I(\nu, n, \alpha(t)) = \big(n\nu + \alpha(t)\big) \dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) \\ - n\nu \dot{F}^{-1}(\nu) - nF\big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\big) + nF\big(\dot{F}^{-1}(\nu)\big). \end{aligned} \tag{2.11}$$

Proposition 1. *The selected arm at each time t for the BMLE algorithm under exponential family*

rewards is

$$\pi_t^{BMLE} = \underset{i\in\{1,\cdots,N\}}{\operatorname{argmax}}\ I(p_i(t), N_i(t), \alpha(t)). \tag{2.12}$$

The proof is in Appendix A.5. To further substantiate the above index, we examine closed-form expressions of the BMLE indices for three commonly-studied distributions.

2.4.2.1 Bernoulli Distributions

For Bernoulli distributions, we know $F(\eta) = \log(1+e^{\eta})$, $\dot{F}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$, $\dot{F}^{-1}(\theta) = \log(\frac{\theta}{1-\theta})$, and $F(\dot{F}^{-1}(\theta)) = \log(\frac{1}{1-\theta})$. Based on (2.11), the BMLE index derived from the Bernoulli rewards can be obtained as follows. We define $\tilde{p}_i(t) := \min\{p_i(t) + \alpha(t)/N_i(t), 1\}$.

Corollary 1. *For the Bernoulli rewards, the BMLE index given by (2.11) becomes*

$$I(p_i(t), N_i(t), \alpha(t)) \tag{2.13}$$

$$= N_i(t)\big\{\tilde{p}_i(t)\log\tilde{p}_i(t) + (1-\tilde{p}_i(t))\log(1-\tilde{p}_i(t)) \tag{2.14}$$

$$- p_i(t)\log(p_i(t)) - (1-p_i(t))\log(1-p_i(t))\big\}. \tag{2.15}$$

The detailed proof of Corollary 1 is provided in Appendix A.6. From Corollary 1, we observe that BMLE is an index-type algorithm with index $I(p_i(t), N_i(t), \alpha(t))$ for arm i that is easy to compute.

Remark 1. The index in (2.13)-(2.15) can be reorganized as

$$I(p_i(t), N_i(t), \alpha(t)) \tag{2.16}$$

$$= \alpha(t)\log\frac{\tilde{p}_i(t)}{1-\tilde{p}_i(t)} - N_i(t)\cdot \mathrm{KL}(p_i(t)\,||\,\tilde{p}_i(t)), \tag{2.17}$$

where $\mathrm{KL}(\theta'\,||\,\theta'')$ denotes the Kullback–Leibler divergence between a Bernoulli(θ') and Bernoulli(θ'') distribution. The derivation of this index is provided in Appendix A.7. Through this alternative expression, one may find some connection with the KL-UCB algorithm [4, 5, 7, 6], which selects the arm with index: $\operatorname{argmax}_i \max\{q\in[0,1] : N_i(t)\cdot \mathrm{KL}(p_i(t)\,||\,q) \le \log t + 3\log\log t\}$. The index

of (2.17) however, has two salient distinctions: (i) The BMLE index is derived from the machinery of maximum likelihood estimation, while KL-UCB originates from the idea of introducing more smoothness into the UCB-type algorithms. (ii) Instead of solving a convex optimization problem for obtaining the index as KL-UCB, the BMLE index enjoys a simple closed-form expression.

Remark 2. The expression for $\tilde{p}_i(t)$ resembles that of a Bayes estimator (under quadratic loss) for a Binomial likelihood with an improper Beta prior to the success probability. However, BMLE is not a Bayesian approach as it does not impose any prior distribution on the model parameters. Instead, BMLE achieves exploration entirely through the time-varying bias term.

2.4.2.2 Gaussian Distributions

For Gaussian reward distributions with the same variance σ^2 among arms, we know $F(\eta_i) = \sigma^2\eta_i^2/2$, $\dot{F}(\eta_i) = \sigma^2\eta_i$, $\dot{F}^{-1}(\theta_i) = \theta_i/\sigma^2$, and $F(\dot{F}^{-1}(\theta_i)) = \theta_i^2/2\sigma^2$, for each arm i. Based on (2.11), the BMLE index for the Gaussian rewards can be derived as follows.

Corollary 2. *For Gaussian reward distributions with the same variance σ^2 among arms, under the BMLE algorithm, the selected arm at each time t is*

$$\pi_t^{BMLE} = \underset{i\in\{1,\cdots,N\}}{\operatorname{argmax}} \Big\{p_i(t) + \frac{\alpha(t)}{2N_i(t)}\Big\}. \tag{2.18}$$

The proof of Corollary 2 is provided in Appendix A.8.

Remark 3. The index in (2.18) has a similar flavor to UCB-type indices [1, 2, 7]. However, it is directly derived from the machinery of maximum likelihood estimation without resorting to concentration inequalities.

2.4.2.3 Exponential Distributions

Corollary 3. *For exponential reward distributions, under the BMLE algorithm, the selected arm at each time t is*

$$\pi_t^{BMLE} = \underset{i\in\{1,\cdots,N\}}{\operatorname{argmax}} \big\{N_i(t)\log\big(\frac{N_i(t)p_i(t)}{N_i(t)p_i(t)+\alpha(t)}\big)\big\}. \tag{2.19}$$

The proof of Corollary 3 is provided in Appendix A.9.

Remark 4. While the BMLE indices are derived from parametric distributions, they can be readily applied to other *non-parametric* reward distributions. As will be seen in Proposition 4 and Proposition 5, in such *misspecified* settings, the derived indices still achieve logarithmic regret bounds for non-parametric distributions that satisfy some concentration inequalities.

2.4.3 Properties of the Derived BMLE Index

We introduce several useful properties of the index $I(\nu, n, \alpha(t))$ in (2.11) to better demonstrate the behavior of the proposed BMLE algorithm. To begin with, we discuss the dependence of $I(\nu, n, \alpha(t))$ on ν and n.

Lemma 1. *For a fixed $\nu \in \Theta$ and $\alpha(t) > 0$, $I(\nu, n, \alpha(t))$ is strictly decreasing with n, for all $n > 0$.*

Lemma 2. *For a fixed $n > 0$ and $\alpha(t) > 0$, $I(\nu, n, \alpha(t))$ is strictly increasing with ν, for all $\nu \in \Theta$.*

The proofs of Lemmas 1 and 2 are provided in Appendices A.1 and A.2. Recall that the BMLE index is $I(p_i(t), N_i(t), \alpha(t))$, where $p_i(t)$ denotes the empirical mean. Then, it is reasonable that the index of an arm increases with its empirical mean reward, as suggested by Lemma 2.

To prepare for the following lemmas, we first define a function $\xi(k; \nu) : \mathbb{R}_{++} \to \mathbb{R}$ as

$$\xi(k;\nu) = k\Big[\big(\nu + \frac{1}{k}\big)\dot{F}^{-1}(\nu + \frac{1}{k}) - \nu\dot{F}^{-1}(\nu)\Big] \tag{2.20}$$

$$- k\Big[F\Big(\dot{F}^{-1}\big(\nu + \frac{1}{k}\big)\Big) - F(\dot{F}^{-1}(\nu))\Big]. \tag{2.21}$$

It is easy to verify that $I(\nu, k\alpha(t), \alpha(t)) = \alpha(t)\xi(k; \nu)$. By Lemma 1, we know $\xi(k; \nu)$ is strictly decreasing with k. Moreover, define a function $K^*(\theta', \theta'')$ as

$$K^*(\theta', \theta'') = \inf\{k : \dot{F}^{-1}(\theta') > \xi(k; \theta'')\}. \tag{2.22}$$

Lemma 3. *Given any pair of real numbers $\mu_1, \mu_2 \in \Theta$ with $\mu_1 > \mu_2$, for any real numbers n_1, n_2 that satisfy $n_1 > 0$ and $n_2 > K^*(\mu_1, \mu_2)\alpha(t)$ (with $K^*(\mu_1, \mu_2)$ being finite), we have $I(\mu_1, n_1, \alpha(t)) > I(\mu_2, n_2, \alpha(t))$.*

Lemma 4. *Given any real numbers $\mu_0, \mu_1, \mu_2 \in \Theta$ with $\mu_0 > \mu_1$ and $\mu_0 > \mu_2$, for any real numbers n_1, n_2 that satisfy $n_1 \leq K^*(\mu_0, \mu_1)\alpha(t)$ and $n_2 > K^*(\mu_0, \mu_2)\alpha(t)$, we have $I(\mu_1, n_1, \alpha(t)) > I(\mu_2, n_2, \alpha(t))$.*

The proofs of Lemmas 3 and 4 are in Appendices A.3 and A.4. Note that Lemma 3 shows that BMLE indeed tends to avoid the arm with a smaller empirical mean reward after sufficient exploration which is quantified in terms of $\alpha(t)$ by $n_2 > K^*(\mu_1, \mu_2)\alpha(t)$. On the other hand, Lemma 4 suggests that BMLE is designed to continue exploration even if the empirical mean reward is initially fairly low (which is reflected by the fact that there is no restriction on the ordering between μ_1 and μ_2 in Lemma 4), when there has been insufficient exploration, as quantified by $n_1 \leq K^*(\mu_0, \mu_1)\alpha(t)$.

2.5 Regret Analysis of the BMLE Algorithm

In this section, we present a theoretical analysis of the proposed bandit algorithm.

2.5.1 Exponential Families With a Lower Bound on Mean

We consider the regret performance of BMLE for the exponential families with a known lower bound on the mean (denoted by $\underline{\theta}$). For example, the mean of an exponential distribution is non-negative and, therefore $\underline{\theta} = 0$. Note that such a collection naturally includes the commonly-studied exponential families that are defined on the positive half real line, such as the exponential, Binomial, Poisson, and Gamma (with a fixed shape parameter).

Proposition 2. *For any exponential family with a lower bound $\underline{\theta}$ on the mean, for any $\varepsilon \in (0, 1)$, the regret of BMLE using (2.11) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \geq 4/(D(\theta_1 - \frac{\varepsilon\Delta}{2}, \theta_1)K^*(\theta_1 - \frac{\varepsilon\Delta}{2}, \underline{\theta}))$*

satisfies

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[\max \big\{ \frac{4}{D(\theta_a + \frac{\varepsilon\Delta_a}{2}, \theta_a)}, \tag{2.23}$$

$$C_\alpha K^*(\theta_1 - \frac{\varepsilon\Delta_a}{2}, \theta_1 + \frac{\varepsilon\Delta_a}{2}) \big\} \log T + 1 + \frac{\pi^2}{3} \big]. \tag{2.24}$$

Below is a sketch of our proof. Our target is to quantify the expected number of trials of each sub-optimal arm a up to time T. The regret bound proof starts with a similar demonstration as for UCB1 [1] by studying the probability of the event $\{I(p_1(t), N_1(t), \alpha(t)) \leq I(p_a(t), N_a(t), \alpha(t))\}$, using the Chernoff bound for exponential families. However, it is significantly different from the original proof as the dependency between the level of exploration, and the bias term $\alpha(t)$ is technically more complex, compared to the straightforward confidence interval used by the conventional UCB-type policies. Specifically, the main challenge lies in characterizing the behavior of the BMLE index for both regimes where $N_1(t)$ is small compared to $\alpha(t)$, as well as when it is large compared to $\alpha(t)$. Such a challenge is handled by considering three cases separately: (i) Consider $N_1(t) > \frac{4}{D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1)} \log t$ and apply Lemma 3; (ii) Consider $N_1(t) \leq \frac{4}{D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1)} \log t$ and $N_1(t) \leq K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta})\alpha(t)$ and apply Lemma 4; (iii) Use Lemma 4 to show that $\{N_1(t) \leq \frac{4}{D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1)} \log t\}$ and $\{N_1(t) > K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta})\alpha(t)\}$ cannot occur simultaneously. The complete proof is provided in Appendix A.10.

2.5.2 Gaussian Distributions

Proposition 3. *For Gaussian reward distributions with variance bounded by σ^2 for all arms, the regret of BMLE using (2.18) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \geq \frac{256\sigma^2}{\Delta}$ satisfies*

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[\frac{2}{\Delta_a} C_\alpha \log T + \frac{2\pi^2}{3} \big]. \tag{2.25}$$

Below is a sketch of our proof. We extend the proof procedure of Proposition 2 for Gaussian rewards, with the help of Hoeffding's inequality. We then prove an additional lemma, which shows that conditioned on the "good" events, the BMLE index of the optimal arm (i.e. arm 1) is always

larger than that of a sub-optimal arm a if $N_a(t) \geq \frac{2}{\Delta_a}\alpha(t)$ and $\alpha(t) \geq \frac{256\sigma^2}{\Delta_a}$, regardless of $N_1(t)$. The complete proof is provided in Appendix A.11.

2.5.3 Beyond Parametric Distributions

As mentioned in Remark 4, BMLE indices derived for exponential families can be readily applied to other non-parametric distributions. Moreover, the regret proofs in Propositions 2-3 can be readily extended if the non-parametric rewards also satisfy proper concentration inequalities. Below we define two classes of reward distributions, namely sub-Gaussian and sub-exponential [51].

Definition 1. *A random variable X with mean $\mu = \mathbb{E}[X]$ is σ-sub-Gaussian if there exists $\sigma > 0$ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}}, \ \forall \lambda \in \mathbb{R}. \tag{2.26}$$

Definition 2. *A random variable X with mean $\mu = \mathbb{E}[X]$ is (ρ, κ)-sub-exponential if there exist $\rho, \kappa \geq 0$ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\rho^2\lambda^2}{2}}, \ \forall |\lambda| < \frac{1}{\kappa}. \tag{2.27}$$

Proposition 4. For any σ-sub-Gaussian reward distributions, BMLE using (2.18) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \geq \frac{256\sigma^2}{\Delta}$ yields $\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[\frac{2}{\Delta_a} C_\alpha \log T + \frac{2\pi^2}{3}\big]$.

The proof of Proposition 3 still holds for Proposition 4 without any change as Hoeffding's inequality directly works for sub-Gaussian distributions.

Proposition 5. For any (ρ, κ)-sub-exponential reward distributions defined on the positive half line, BMLE using (2.18) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \geq 16(\kappa\varepsilon\Delta + 2\rho^2)/((\varepsilon\Delta)^2 K^*(\theta_1 - \frac{\varepsilon\Delta}{2}, 0))$ achieves a regret bound

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[1 + \frac{\pi^2}{3} + \max\big\{\frac{16(\kappa\varepsilon\Delta + 2\rho^2)}{(\varepsilon\Delta_a)^2}, \tag{2.28}$$

$$C_\alpha K^*(\theta_1 - \frac{\varepsilon\Delta_a}{2}, \theta_1 + \frac{\varepsilon\Delta_a}{2})\big\} \log T\big]. \tag{2.29}$$

The proof of Proposition 2 can be easily extended for Proposition 4 by replacing the Chernoff

bound with the sub-exponential tail bounds. The proof of Proposition 5 is provided in Appendix A.12.

2.6 Empirical Study on the Performance of the BMLE Algorithm

To evaluate the effectiveness and efficiency of BMLE, we conducted a comprehensive empirical comparison between BMLE and other methods for Bernoulli bandits, Gaussian bandits, and exponential bandits. We paid particular attention to the fairness of the comparison and reproducibility of the experimental results. To ensure the sample path is the same for all methods in each round of decision-making, we prepared data containing the outcomes of pulling all arms over all rounds in advance of each experiment. As such, in each round, the outcome of pulling one arm can be obtained directly through querying the prepared data, instead of calling a random generator. We also note that a few benchmark methods such as Thompson Sampling and sample-based IDS/VIDS will change the state of the underlying random generator and thus influence each other's sample path when compared together in one program, thus bringing unfairness into comparison. This is because, in each round, they need to sample from random generators based on updated posteriors. To avoid this unfairness to occur, we evaluate their performance separately with the same prepared data, and the same seed for the random number generators, i.e., the calling of random generators in one method will not change the state of random generators in other methods. To ensure the reproducibility of experimental results, we set up the seeds for the random number generators at the beginning of each experiment and provide all the codes, including the seed setup in GitHub.

2.6.1 An Adaptive Scheme for Selecting Bias in BMLE

As discussed in Section 2.5, BMLE achieves logarithmic regret by choosing $\alpha(t) = C_\alpha \log t$, where the choice of C_α involves the minimum gap Δ and the largest mean θ_1. We consider the following adaptive scheme that gradually learns Δ and θ_1. To illustrate the overall procedure, we use the C_α in Proposition 5 as an example (for ease of notation, we use $C_{\alpha,0}$ to denote the constant $16(\kappa\varepsilon\Delta + 2\rho^2)/((\varepsilon\Delta)^2 K^*(\theta_1 - \frac{\varepsilon\Delta}{2}, 0))$).

- **Estimate Δ and θ_1**: Note that Δ can be expressed as $\max_{1\le i\le N}\{\theta_i - \max_{j\neq i}\theta_j\}$. For each arm i,

construct $U_i(t)$ and $L_i(t)$ as the upper and lower confidence bounds of $p_i(t)$ based on proper concentration inequalities. Then, construct an estimator of Δ as $\hat{\Delta}_t := \max_{1\le i\le N}\big\{\max\big(0, L_i(t) - \max_{j\neq i} U_j(t)\big)\big\}$. Meanwhile, we use $U_{\max}(t) := \max_{1\le i\le N} U_i(t)$ as an estimate of θ_1. Based on the confidence bounds, we know $\hat{\Delta}_t \le \Delta$ and $U_{\max}(t) \ge \theta_1$, with high probability.

- **Construct the bias using estimators**: We construct $\alpha(t) = \min\{\widehat{C}_\alpha(t), \beta(t)\}\log t$, where $\widehat{C}_\alpha(t)$ estimates $C_{\alpha,0}$ by replacing Δ with $\hat{\Delta}_t$ and θ_1 with $U_{\max}(t)$, and $\beta(t)$ is a non-negative strictly increasing function satisfying $\lim_{t\to\infty}\beta(t) = \infty$. With high probability, $\widehat{C}_\alpha(t)$ gradually approaches the target value $C_{\alpha,0}$ from above as time evolves. On the other hand, $\beta(t)$ guarantees smooth exploration initially and will ultimately exceed $\widehat{C}_\alpha(t)$.

2.6.2 Pseudo Code of the Adaptive Scheme

In this section, we provide the pseudo-code of the experiments in Section 2.6. To begin with, Algorithm 1 shows the pseudo-code for choosing the bias term $\alpha(t)$ in Bernoulli bandits. The main idea is to learn a proper C_α considered in the regret analysis by gradually increasing C_α until it is sufficiently large. This is accomplished by setting $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\}\log t$ (Line 13 in Algorithm 1), where $\beta(t)$ is a positive strictly increasing function with $\lim_{t\to\infty}\beta(t) = \infty$ (e.g. $\beta(t) = \sqrt{\log t}$ in the experiments in Section 2.6), and $\hat{C}_\alpha(t)$ serves as an over-estimate of the minimum required C_α based on the estimators $\hat{\Delta}_t$ and $U_{\max}(t)$ for Δ and θ_1 (Lines 3-8 in Algorithm 1). Note that Δ can be written as $\Delta = \max_{1\le i\le N}\{\theta_i - \max_{j\neq i}\theta_j\}$. Therefore, $\hat{\Delta}_t$ is a conservative estimate of Δ in the sense that $\hat{\Delta}_t \le \Delta$, conditioned on the high probability events $\theta_i \in [L_i(t), U_i(t)]$, for all i. Here the confidence bounds $L_i(t)$ and $U_i(t)$ are constructed with the help of Hoeffding's inequality. For small t, it is expected that $\hat{\Delta}_t$ is very close to zero and hence $\hat{C}_\alpha(t)$ is large. Therefore, initially $\beta(t)$ serves to gradually increase C_α and guarantees enough exploration after $\beta(t)$ exceeds the minimum required C_α. Given sufficient exploration enabled by $\beta(t)$, the estimate $\hat{\Delta}_t$ gets rather accurate (i.e. $\hat{\Delta}_t \approx \Delta$), and subsequently $\hat{C}_\alpha(t)$ shall be clamped at some value slightly larger than the minimum required C_α. On the other hand, as the calculation of $\hat{C}_\alpha(t)$ involves the subroutine of searching for the value $K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)$, we can accelerate

Algorithm 1 Adaptive Scheme for Choosing $\alpha(t)$ in Bernoulli Bandits

1: **Input:** N, $\varepsilon \in (0, \frac{1}{2})$, and $\beta(t)$
2: **for** $t = 1, 2, \cdots$ **do**
3: **for** $i = 1$ **to** N **do**
4: $U_i(t) = \min\Big(p_i(t) + \sqrt{(N+2)\log t / N_i(t)}, 1\Big)$ // upper confidence bound of the empirical mean
5: $L_i(t) = \max\Big(p_i(t) - \sqrt{(N+2)\log t / N_i(t)}, 0\Big)$ // lower confidence bound of the empirical mean
6: **end for**
7: $U_{\max}(t) = \max_{i=1,\cdots,N} U_i(t)$
8: $\hat{\Delta}_t = \max_i \Big\{ \max\big(0, L_i(t) - \max_{j \neq i} U_j(t)\big) \Big\}$
9: **if** $\xi\big(\frac{N+2}{2(\varepsilon\hat{\Delta}_t)^2\beta(t)}, 0\big) < \dot{F}^{-1}(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2})$ **then**
10: $\alpha(t) = \beta(t)\log t$ // In this case, we know $\hat{C}_\alpha(t) > \beta(t)$
11: **else**
12: Find $\hat{C}_\alpha(t) = \frac{N+2}{2(\varepsilon\hat{\Delta}_t)^2 K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)}$ by solving the minimization problem of (2.22) for $K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)$.
13: $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\}\log t$
14: **end if**
15: **end for**

the adaptive scheme by first checking if it is possible to have $\hat{C}_\alpha(t) \geq \beta(t)$. Equivalently, this can be done by quickly verifying whether $\xi(\frac{N+2}{2(\varepsilon\hat{\Delta}_t)^2\beta(t)}, 0) < \dot{F}^{-1}(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2})$ (Line 9 in Algorithm 1).

Similarly, Algorithms 2 and 3 demonstrate the pseudo codes for selecting $\alpha(t)$ in exponential bandits and Gaussian bandits, respectively. Compared to the Bernoulli case, the main difference of the exponential case lies in the construction of the confidence bounds (Lines 4-5 in Algorithm 2), which leverage the sub-exponential tail bounds instead of Hoeffding's inequality. On the other hand, Algorithm 3 for the Gaussian case differs from the other two in that it does not require the calculation of $K^*(\cdot, \cdot)$ and only $\hat{\Delta}_t$ is needed (Lines 7-8 in Algorithm 3).

2.6.3 Detailed Description of the Major Competitors

The competitors from the frequentist setting include UCB [1], UCB-Tuned (UCBT) [1], MOSS [2, 3], and KL-UCB [5]. On the other hand, the competitors from Bayesian family consist of Knowl-

Algorithm 2 Adaptive Scheme for Choosing $\alpha(t)$ in Exponential Bandits

Input: N, $\varepsilon \in (0, \frac{1}{2})$, and $\beta(t)$
for $t = 1, 2, \cdots$ **do**
  for $i = 1$ **to** N **do**
    $U_i(t) = p_i(t) + \frac{\kappa(N+2)\log t + \sqrt{\kappa^2(N+2)^2(\log t)^2 + 2\rho^2(N+2)\log t}}{N_i(t)}$ // upper confidence bound
    $L_i(t) = \max\left(p_i(t) - \frac{\kappa(N+2)\log t + \sqrt{\kappa^2(N+2)^2(\log t)^2 + 2\rho^2(N+2)\log t}}{N_i(t)}, 0\right)$ // lower confidence bound
  end for
  $U_{\max}(t) = \max_{i=1,\cdots,N} U_i(t)$
  $\hat{\Delta}_t = \max_i \left\{ \max\left(0, L_i(t) - \max_{j \neq i} U_j(t)\right) \right\}$
  if $\xi\left(\frac{16(\kappa\varepsilon\hat{\Delta}_t + 2\rho^2)}{(\varepsilon\hat{\Delta}_t)^2\beta(t)}, 0\right) < \dot{F}^{-1}(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2})$ **then**
    $\alpha(t) = \beta(t)\log t$ // In this case, we know $\hat{C}_\alpha(t) > \beta(t)$
  else
    Find $\hat{C}_\alpha(t) = \frac{16(\kappa\varepsilon\hat{\Delta}_t + 2\rho^2)}{(\varepsilon\hat{\Delta}_t)^2 K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)}$ by solving the minimization problem of (2.22) for $K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)$. $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\}\log t$
  end if
end for

Algorithm 3 Adaptive Scheme for Choosing $\alpha(t)$ in Gaussian Bandits

Input: N, σ, and $\beta(t)$
for $t = 1, 2, \cdots$ **do**
  for $i = 1$ **to** N **do**
    $U_i(t) = p_i(t) + \sqrt{(N+2)\log t / N_i(t)}$ // upper confidence bound of the empirical mean
    $L_i(t) = p_i(t) - \sqrt{(N+2)\log t / N_i(t)}$ // lower confidence bound of the empirical mean
  end for
  $\hat{\Delta}_t = \max_i \left\{ \max\left(0, L_i(t) - \max_{j \neq i} U_j(t)\right) \right\}$
  Calculate $\hat{C}_\alpha(t) = \frac{256\sigma^2}{\hat{\Delta}_t}$
  $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\}\log t$
end for

edge Gradient (KG) [12], its variant - KG* [13], and the approximation KG* - KG(min) (KGMin), and MN (KGMN) [31] in [31]), Thompson sampling (TS) [8, 10, 52], Bayes-UCB (BUCB) [7], Information Directed Sampling (IDS) and its variant - variance-based IDS (VIDS) [14, 15], and GPUCB [29] and its tuned version - GPUCB-Tuned (GPUCBT). GPUCB and GPUCBT are only

regarded as competitors of BMLE in Gaussian bandits.

2.6.3.1 Frequentist Approaches

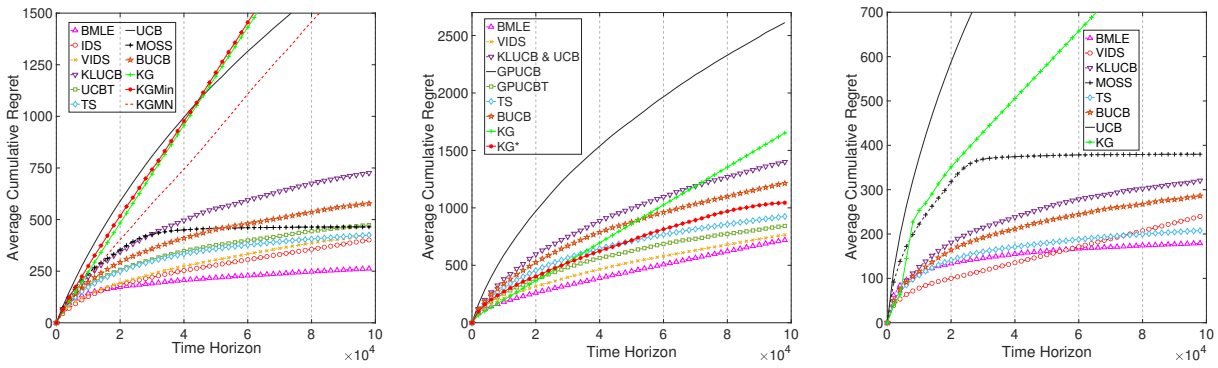
The UCB algorithm selects an arm i which maximizes the index $\widehat{\theta}_i(t) + \sqrt{2\log(t)/N_i(t)}$, where $\widehat{\theta}_i(t)$ is the empirical mean reward received from samples of arm i. The index of UCB is constructed to facilitate regret bound analysis. Its empirical performance is often unsatisfactory. To achieve better empirical performance, UCB-Tuned (UCBT) replaces the index by $\widehat{\theta}_i(t) + \sqrt{\min\{1/4, \overline{V}_t(i)\}\log(t)/N_i(t)\}}$, where $\overline{V}_t(i)$ is the upper bound on the variance of the reward of arm i. UCBT often demonstrates outstanding empirical performance, but unfortunately, there is limited literature illustrating any guarantee to its regret bound. The MOSS algorithm uses $p_i(t) + \sqrt{\max(\log(\frac{T}{N_i(t)\cdot N}), 0)/N_i(t)}$, a slightly different index from UCB and UCBT for arm i, where the computation of the index requires additional knowledge of time horizon T. It automatically decreases the amount of exploration after an arm has already been pulled more than T/N times. It has two limitations: (1) sub-optimality - it is only *nearly* asymptotically optimal, and can be arbitrarily worse than UCB in some regimes, and (2) instability - the distribution of its regret can be not well-behaved. KL-UCB is currently the most computationally heavy method in the frequentist family. Taking Bernoulli bandits as an example, the index for arm i is obtained through solving an optimization problem $\max\{p \in [0,1] : D(p_i(t), p) \leq (\log(t) + c\log(\log(t))/N_i(t)\}$. The optimization problem often relies on Newton's method or bisection search for its solution, except in Gaussian bandits, where a closed-form index can be derived thanks to the tractable form of KL divergence for Gaussian distributions. KL-UCB often demonstrates better empirical performance than UCB as it constructs the upper confidence bound using Chernoff's inequality, a tighter one than that which has been used in UCB.

2.6.3.2 Bayesian Approaches

GPUCB and GPUCBT are only for Gaussian bandits. Under GPUCB, the index of arm i at time t is $\mu_i(t) + \sqrt{\beta_t}\sigma_i(t)$, where $\mu_i(t)$ and $\sigma_i(t)$ are the posterior mean and posterior standard deviation for arm i and $\beta_t = 2\log(Nt^2\pi^2/6\delta)$. They provide regret bounds that hold with probability at

least $1-\delta$. As such, δ is often chosen to be a very small positive number close to zero [15]. Its variant GPUCB-Tuned (GPUCBT) demonstrates better empirical performance and replaces the original β_t by $\beta_t = c\log(t)$, where c is a hyperparameter tuned for time horizon T. Similar to UCBT, there is also concern about the theoretical guarantee of GPUCBT. Different from UCB-like algorithms, KG is inspired by the Bellman equation and is one type of the one-step look-ahead policy: the index for arm i is determined by an immediate reward of pulling arm i and the expected future rewards after observing the outcomes of the pull. The future rewards are quantified by the knowledge improvement of the optimal arm after observing the outcome of the pulled arm in the current round. To be more specific, KG uses the index $\mu_i(t) + \mathbb{E}[\mu_*(t+1) - \mu_*(t)|i]$ for arm i, where $\mu_*(t) = \max_i\{\mu_i(t)\}$. KG has a closed-form in both Bernoulli and Gaussian bandits. However, beyond limitations such as requiring specification of the time horizon T, the regret of KG sometimes grows linearly as it may explore insufficiently, especially when the outcome of true distribution is discrete and the time horizon is long [31]. To overcome this limitation, KG* was proposed by extending the one-step look-ahead to multi-step look-ahead. At time t, KG* calculates the index for an arm over all possible steps of look-ahead and thus suffers from high computational complexity, and scales very poorly with time horizon T. To enable longer time horizons, the heuristic approximation methods KGMin and KGMN were applied [31]. Basically, they use the golden section search to approximately maximize a non-concave function but are still empirically effective, as illustrated in [15]. TS, in general, works well for different types of bandits, including those with discrete and continuous outcomes. It often outperforms vanilla UCB algorithms and follows a simple intuition: select the arm according to the probability that it is the optimal arm. In each round, TS uses the outcome drawn from the posterior distribution of arm i as its index for arm i. Motivated by the Bayesian interpretation of the problem and to retain the simplicity of UCB-like algorithms in implementation, BUCB constructs upper confidence bounds based on the $1-1/t$ quantiles of the posterior distribution. Specially, BUCB uses $Q_i\Big(1-1/\big(t(\log(T))^c\big)\Big)$ as the index for arm i, where $Q_i\big(\,\cdot\,\big)$ denotes the quantile function of the posterior distribution for arm i. Like MOSS and GPUCBT, BUCB also requires knowledge of time horizon T. TS and

BUCB are proved to be optimal, and observed to exhibit excellent performance in experiments with Bernoulli bandits. Their leading positions in empirical performance were displaced by IDS and its variant VIDS. In each round, IDS calculates the probabilities of pulling individual arms through solving an optimization problem $\min_{\boldsymbol{\pi}}\{\Delta_t^2(\boldsymbol{\pi})/g_t(\boldsymbol{\pi})\}$, where $\boldsymbol{\pi}$ is a vector of probabilities and $\boldsymbol{\pi}(i)$ denotes the probability of pulling arm i. $\Delta_t(\boldsymbol{\pi})$ denotes the expected regret under $\boldsymbol{\pi}$ and reward randomness. $g_t(\boldsymbol{\pi})$ denotes the entropy reduction with respect to the optimal arm. The exact computation of $\boldsymbol{\pi}$ and $g_t(\boldsymbol{\pi})$ requires computation of multi-dimensional integrals, which is very expensive. Therefore, usually, sample-based estimation is applied. VIDS further approximates IDS through approximating $g_t(\boldsymbol{\pi})$ by $\boldsymbol{\pi}^\top \boldsymbol{v}$, where $\boldsymbol{v}$ denotes the vector of variance for all arms.



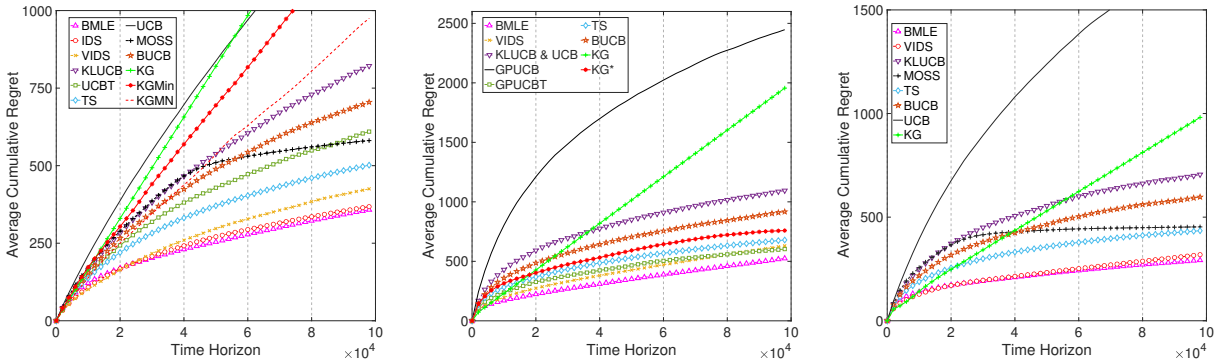
(a) (b) (c)

Figure 2.1: Average cumulative regret over 100 trials with $T = 10^5$ and $\beta(t) = \sqrt{\log(t)}$: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} = (0.66, 0.67, 0.68, 0.69, 0.7, 0.61, 0.62, 0.63, 0.64, 0.65)$; (b) Gaussian bandits with $\sigma = 1$ and $(\mu_i)_{i=1}^{10} = (0.41, 0.52, 0.66, 0.43, 0.58, 0.65, 0.48, 0.67, 0.59, 0.63)$; (c) Exponential bandits with the rates for different arms $(\lambda_i)_{i=1}^{10} = (1/0.31, 1/0.1, 1/0.2, 1/0.32, 1/0.33, 1/0.29, 1/0.2, /0.3, 1/0.15, 1/0.08)$. We use UCBT, GPUCBT and BUCB as the shorthand of UCB-Tuned, GPUCB-Tuned and Bayes-UCB, respectively)

2.6.4 Effectiveness of the BMLE Algorithm

Figures 2.1-2.3 illustrate the comparison of BMLE with major competitors with respect to the average cumulative regret in examples from Bernoulli bandits, Gaussian bandits, and exponential

bandits. In Bernoulli bandits, the reward of an arm i is binary and drawn independently from a Bernoulli distribution with an unknown parameter $\theta_i \in (0, 1)$. In the setting of Gaussian bandits, the reward distribution of arm i is a Gaussian distribution with mean μ_i and the standard deviation σ_i. For ease of presentation and to use the results of Proposition 3, we take $\sigma_i \equiv \sigma$ for all i and assume knowledge of σ in the experiments with Gaussian bandits. In exponential bandits, the reward distribution of arm i follows an exponential distribution with the rate λ_i.

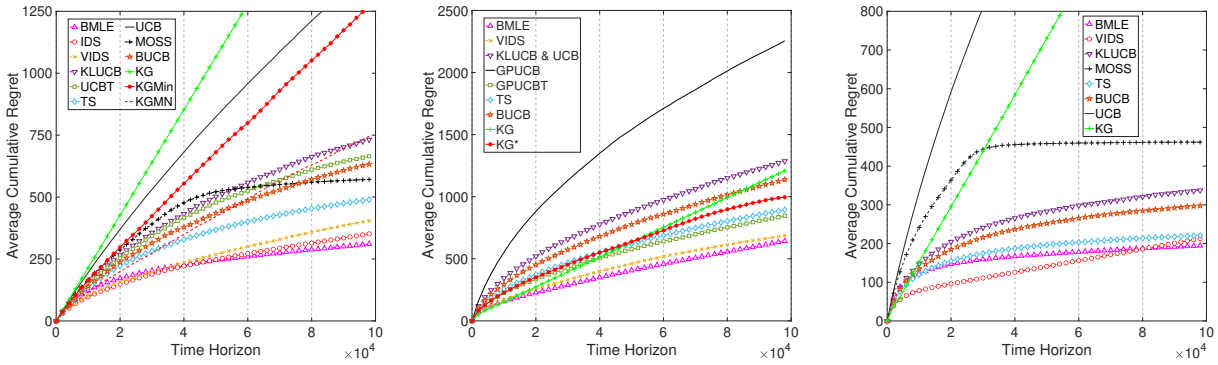


(a) (b) (c)

Figure 2.2: Average cumulative regret over 100 trials with $T = 10^5$ and $\beta(t) = \sqrt{\log(t)}$: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} = (0.655, 0.6, 0.665, 0.67, 0.675, 0.68, 0.685, 0.69, 0.695, 0.7)$; (b) Gaussian bandits with $\sigma = 1$ and $(\mu_i)_{i=1}^{10} = (0.5, 0.75, 0.4, 0.6, 0.55, 0.76, 0.68, 0.41, 0.52, 0.67)$; (c) Exponential bandits with the rates for different arms $(\lambda_i)_{i=1}^{10} = (1/0.46, 1/0.45, 1/0.5, 1/0.48, 1/0.51, 1/0.4, 1/0.43, 1/0.42, 1/0.45, 1/0.44)$.

In the comparison with IDS and VIDS, we sampled 100 points over $[0, 1]$ interval for q (Algorithm 4 in [15]) and $M = 10000$ (Algorithm 3 in [15]). We take $c = 0$ in BUCB and KL-UCB, which are reported to achieve the best empirical performance in the original papers. We take $c = 0.9$ in GPUCBT after parameter tuning as in [15]. In searching for a solution of KL-UCB and computing the value of $C_{\alpha,0}$, the maximum number of iterations is set to be 100. It is also worth mentioning that KL-UCB has the same closed-form index as UCB in Gaussian bandits. The conjugate priors for methods of Bayesian family are Beta distribution $\beta(1, 1)$ for Bernoulli ban-

dits, $\mathcal{N}(0,1)$ for Gaussian bandits, and Gamma distribution $\gamma(1,1)$ for exponential bandits. Most competitors are compared in all three types of bandits. Some are not compared in one or two because their performance is found to be much worse than the rest, e.g., the UCBT for Gaussian and exponential bandits.



(a) (b) (c)

Figure 2.3: Average cumulative regret over 100 trials with $T = 10^5$ and $\beta(t) = \sqrt{\log(t)}$: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} = (0.755, 0.76, 0.765, 0.77, 0.775, 0.78, 0.785, 0.79, 0.795, 0.8)$; (b) Gaussian bandits with $\sigma = 1$ and $(\mu_i)_{i=1}^{10} = (0.65, 0.35, 0.66, 0.4, 0.65, 0.64, 0.55, 0.4, 0.57, 0.54)$; (c) Exponential bandits with the rates for different arms $(\lambda_i)_{i=1}^{10} = (1/0.25, 1/0.28, 1/0.27, 1/0.3, 1/0.29, 1/0.22, 1/0.21, 1/0.24, 1/0.23, 1/0.26)$.

We note that KG performs poorly as it explores insufficiently. This is not surprising as several papers have pointed out the limitations of KG-family methods when rewards are discrete or the time horizon is long [31]. When the rewards are continuous (e.g., in Gaussian bandits and exponential bandits), KG* and the variants of KG* (KGMin, KGMN) achieve remarkable performance improvement over the vanilla KG. This benefits from the consideration of more than one step in look-ahead planning. It is worth emphasizing that BMLE, in general, outperforms all other baselines, including KL-UCB, IDS, VIDS, in terms of regret performance in all three types of bandits. It is not surprising that VIDS or IDS are the closest competitors to BMLE in terms of regret performance. However, BMLE is found to be slightly better than IDS and VIDS in those examples.

Moreover, in spite of the good performance of IDS and VIDS, the determination of their indices suffers from high computational complexity, even under sample-based approximation. In contrast, the BMLE index, with its simple closed-form expression is trivial to compute. One more advantage of BMLE over some of the baselines is that it is "time horizon agnostic", i.e., the computation of the BMLE index does not need the knowledge of time horizon T. In contrast, BUCB, MOSS, GPUCBT, and KG-family methods (KG, KG*) need to know T. It needs to be emphasized that we evaluate the effectiveness of BMLE in both challenging examples as well as randomly picked examples. For instance, for Gaussian bandits, we choose the parameter values to make the problem very challenging: the standard deviation is 100 times the value difference between the largest mean and second-largest mean. In contrast, the means of exponential bandits are chosen more randomly and easier to be differentiated.

Tables 2.2-2.10 provide detailed statistics, including the mean as well as the standard error and quantiles of the final regrets, with the row-wise smallest values highlighted in boldface. From the tables, we observe that BMLE tends to have the smallest value of regret at medium to high quantiles, and comparable to the smallest values at other lower quantiles among those that have comparable mean values (e.g., IDS, VIDS, KLUCB). Along with the presented statistic of standard error, they suggest that the BMLE's performance enjoys comparable robustness as those baselines that achieve similar mean regret.

2.6.5 Efficiency of the BMLE Algorithm

Figures 2.4-2.6 compare the efficiency between BMLE and other baseline methods, where efficiency is represented by the two metrics of interest: Computation time per decision as well as Regret. The computation time per decision is computed through counting the total time spent in each trial and then dividing by #Trials $\times$ T. BMLE is seen to provide promising performance. Especially compared with IDS, VIDS, and KL-UCB, BMLE achieves slightly better regret performance with orders of magnitude less computation time per decision than those methods. IDS and VIDS suffer from high computation complexity because they need to estimate several integrals in each round. The KL-UCB often relies on Newton's method or bisection search to find the index

Stats	BMLE	IDS	VIDS	KLUCB	UCBT	TS	UCB	MOSS	BUCB	KG	KGMin	KGMN
Mean	**2.64**	4.06	4.29	7.30	4.75	4.27	18.1	4.65	5.81	23.8	23.8	18.1
SD	2.34	4.67	4.61	1.09	1.76	1.49	1.13	**0.931**	1.06	21.6	3.55	3.44
Q10	1.42	0.747	1.77	5.84	3.10	2.83	16.5	3.56	4.53	**0.037**	18.99	13.45
Q25	1.61	**1.16**	2.11	6.61	3.63	3.27	17.5	3.94	5.19	10.0	21.0	15.6
Q50	1.91	**1.84**	2.55	7.18	4.49	4.04	18.2	4.59	5.64	20.0	24.1	18.4
Q75	**2.38**	4.61	3.46	8.04	5.44	4.89	18.7	5.21	6.38	40.0	26.2	20.6
Q90	**4.30**	11.4	11.3	8.61	6.55	5.95	19.6	5.71	7.14	50.1	28.1	22.5
Q95	9.93	12.4	12.5	9.26	7.60	6.48	20.0	**6.16**	7.66	69.9	29.1	23.3

Table 2.2: Statistics of distribution of average final regret over 100 trials for the Bernoulli bandits with true values: $(\theta_i)_{i=1}^{10}$ = (0.66, 0.67, 0.68, 0.69, 0.7, 0.61, 0.62, 0.63, 0.64, 0.65) and $T = 10^5$. The regrets are in unit of 100.

Stats	BMLE	IDS	VIDS	KLUCB	UCBT	TS	UCB	MOSS	BUCB	KG	KGMin	KGMN
Mean	**3.62**	3.71	4.30	8.32	6.17	5.06	14.4	5.83	7.16	16.4	13.1	9.91
SD	2.48	2.86	2.69	1.32	1.31	1.56	**0.785**	1.70	1.20	15.9	2.14	1.98
Q10	1.33	1.16	1.97	6.50	4.41	3.34	13.4	4.12	5.64	**0.023**	10.13	7.42
Q25	1.65	**1.64**	2.39	7.32	5.32	3.85	13.9	4.61	6.41	5.02	11.84	8.76
Q50	**2.24**	2.63	3.19	8.23	5.98	4.77	14.4	5.33	7.15	10.0	13.4	9.9
Q75	6.08	**5.69**	5.73	9.30	6.93	5.75	15.0	6.55	7.83	30.0	14.5	11.2
Q90	**6.61**	6.82	8.03	10.2	7.80	6.98	15.4	8.16	8.65	35.0	15.4	12.3
Q95	**7.23**	8.35	9.12	10.6	8.58	7.93	15.6	9.43	9.06	45.0	16.2	13.4

Table 2.3: Statistics of distribution of average final regret over 100 trials for the Bernoulli bandits with true values: $(\theta_i)_{i=1}^{10}$ = (0.655, 0.6, 0.665, 0.67, 0.675, 0.68, 0.685, 0.69, 0.695, 0.7) and $T = 10^5$. The regrets are in unit of 100.

Stats	BMLE	IDS	VIDS	KLUCB	UCBT	TS	UCB	MOSS	BUCB	KG	KGMin	KGMN
Mean	**3.13**	3.56	4.09	7.41	6.70	4.93	14.5	5.72	6.39	21.3	13.0	7.57
SD	2.28	3.87	3.00	1.27	1.20	1.72	**0.691**	1.33	1.28	13.4	2.10	178.3
Q10	1.42	**0.951**	1.67	5.81	5.07	3.29	13.5	4.34	4.86	4.51	10.25	5.31
Q25	1.69	**1.33**	2.01	6.52	5.74	3.74	13.97	4.64	5.43	10.0	11.7	6.29
Q50	2.04	**1.79**	2.68	7.26	6.81	4.63	14.5	5.43	6.23	20.0	13.2	7.54
Q75	**369.7**	5.43	5.56	8.07	7.52	5.34	15.0	6.49	7.24	30.0	14.4	8.97
Q90	**6.80**	6.96	7.88	8.86	8.33	7.26	15.4	7.61	8.14	40.0	15.6	9.64
Q95	**7.2**	8.91	12.0	10.0	8.67	7.74	15.5	8.31	8.67	45.0	15.9	10.4

Table 2.4: Statistics of distribution of average final regret over 100 trials for the Bernoulli bandits with true values: $(\theta_i)_{i=1}^{10}$ = (0.755, 0.76, 0.765, 0.77, 0.775, 0.78, 0.785, 0.79, 0.795, 0.8) and $T = 10^5$. The regrets are in unit of 100.

Stats	BMLE	VIDS	KLUCB&UCB	GPUCB	GPUCBT	TS	BUCB	KG	KG*
Mean	**730.6**	775.0	1412.2	2640.3	848.5	932.7	1222.3	1684.3	1046.0
SD	827.4	678.7	**219.2**	227.0	314.2	282.1	231.4	2056.8	238.9
Q10	135.3	233.9	1147.2	2382.8	529.3	657.8	960.8	**20.4**	788.0
Q25	160.2	336.0	1272.1	2500.0	608.0	706.6	1036.5	**59.9**	891.6
Q50	**263.1**	544.1	1395.9	2600.4	814.7	876.0	1205.9	1035.8	1000.6
Q75	1140.8	1137.7	1545.9	2787.1	**1001.1**	1125.3	1390.6	2028.0	1171.1
Q90	2107.9	1516.5	1674.6	2916.1	**1228.6**	1304.8	1512.9	4028.8	1314.1
Q95	2157.6	1862.0	1724.6	3024.4	1578.7	1472.7	1565.5	7818.3	**1413.7**

Table 2.5: Statistics of distribution of average final regret over 100 trials for the Gaussian bandits with true values: $(\mu_i)_{i=1}^{10} = (0.41, 0.52, 0.66, 0.43, 0.58, 0.65, 0.48, 0.67, 0.59, 0.63)$ and $T = 10^5$.

Stats	BMLE	VIDS	KLUCB&UCB	GPUCB	GPUCBT	TS	BUCB	KG	KG*
Mean	**531.1**	638.5	1102.7	2464.2	607.7	684.3	923.6	1995.0	760.2
SD	469.5	1117.0	196.9	210.8	234.1	250.1	178.7	3541.8	**163.8**
Q10	145.5	143.7	859.7	2200.1	361.4	411.1	724.5	**21.1**	568.4
Q25	167.4	206.6	937.4	2320.7	444.3	501.7	792.9	**30.2**	664.5
Q50	**207.7**	314.1	1093.2	2466.4	544.8	623.1	927.2	1014.4	752.5
Q75	1131.8	889.0	1232.0	2605.0	**714.6**	792.2	1042.0	1044.3	851.4
Q90	1188.1	1183.3	1346.8	2726.0	**926.2**	1058.9	1174.1	8121.5	930.0
Q95	1204.2	1248.6	1439.0	2804.9	1041.8	1209.2	1193.5	9023.5	**959.5**

Table 2.6: Statistics of distribution of average final regret over 100 trials for the Gaussian bandits with true values: $(\mu_i)_{i=1}^{10} = (0.5, 0.75, 0.4, 0.6, 0.55, 0.76, 0.68, 0.41, 0.52, 0.67)$ and $T = 10^5$.

Stats	BMLE	VIDS	KLUCB&UCB	GPUCB	GPUCBT	TS	BUCB	KG	KG*
Mean	**652.0**	694.7	1302.0	2281.0	856.5	903.4	1149.5	1233.6	1001.7
SD	581.8	776.1	**164.5**	169.5	255.8	268.2	201.0	1659.2	234.8
Q10	127.3	193.6	11000.0	2062.5	561.1	574.8	897.0	**24.5**	747.2
Q25	155.7	322.9	1173.4	2156.6	665.7	715.8	1000.4	**72.0**	827.9
Q50	**265.4**	471.9	1295.7	2262.7	814.3	849.3	1130.5	1021.1	944.4
Q75	1116.2	**861.0**	1428.3	2397.7	1007.8	1085.6	1294.0	1987.0	1128.1
Q90	1202.8	1236.1	1492.8	2506.3	**1164.6**	1283.0	1404.6	2028.1	1346.7
Q95	2021.8	1467.5	1549.4	2545.1	**1334.9**	1394.5	1511.5	2055.5	1467.2

Table 2.7: Statistics of distribution of average final regret over 100 trials for the Gaussian bandits with true values: $(\mu_i)_{i=1}^{10} = (0.65, 0.35, 0.66, 0.4, 0.65, 0.64, 0.55, 0.4, 0.57, 0.54)$ and $T = 10^5$.

Stats	BMLE	VIDS	KLUCB	TS	UCB	MOSS	BUCB	KG
Mean	**179.6**	243.3	322.7	208.6	1504.6	379.9	288.2	961.6
SD	119.4	463.1	63.9	61.3	66.1	**44.5**	71.9	1063.3
Q10	128.7	37.6	239.4	132.8	1430.9	329.4	196.7	**26.5**
Q25	139.7	47.9	271.3	157.7	1452.0	345.8	238.3	**37.2**
Q50	155.2	**70.5**	331.7	202.3	1505.4	380.1	275.1	387.2
Q75	173.4	**103.7**	367.2	243.4	1550.6	405.9	330.6	2450.7
Q90	**195.4**	1039.9	407.0	303.1	1586.5	435.0	377.3	2509.9
Q95	**291.7**	1074.1	423.2	320.1	1617.6	457.8	405.3	2522.7

Table 2.8: Statistics of distribution of average final regret over 100 trials for the Exponential bandits with true values: $(1/\lambda_i)_{i=1}^{10}$ = (0.31, 0.1, 0.2, 0.32, 0.33, 0.29, 0.2, 0.3, 0.15,0.08) and $T = 10^5$.

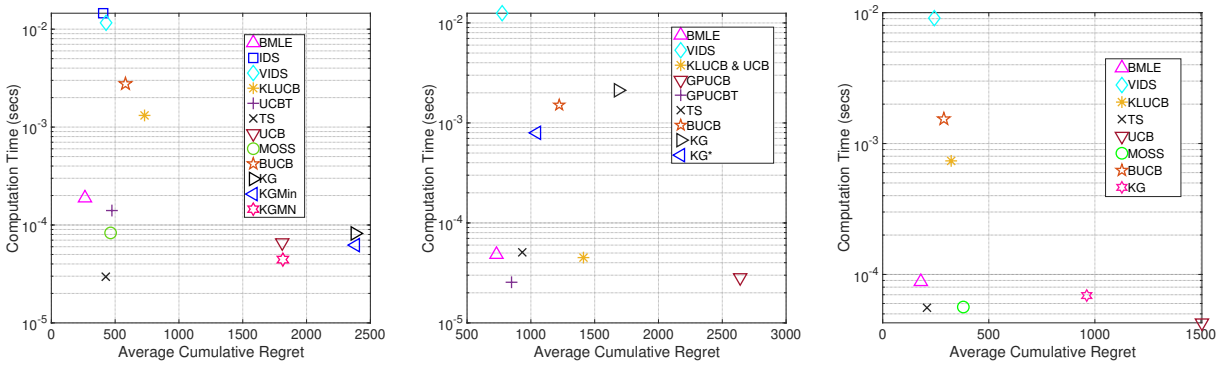
Stats	BMLE	VIDS	KLUCB	TS	UCB	MOSS	BUCB	KG
Mean	**294.6**	322.4	710.6	436.7	1805.6	453.5	600.8	1000.0
SD	301.3	352.5	**118.0**	168.7	126.6	147.8	126.3	1637.9
Q10	139.8	93.3	565.4	288.1	1653.3	342.8	464.1	**34.9**
Q25	148.7	116.4	609.8	335.9	1713.9	374.8	792.9	**30.2**
Q50	176.9	166.1	695.1	411.0	1789.0	419.6	592.2	**77.4**
Q75	**237.4**	273.8	784.9	468.3	1898.1	483.9	662.3	1050.0
Q90	919.0	1064.9	875.6	610.0	1970.0	**578.0**	739.0	4920.6
Q95	1183.3	1112.1	916.6	682.5	2035.3	**644.9**	789.5	5042.0

Table 2.9: Statistics of distribution of average final regret over 100 trials for the Exponential bandits with true values: $(1/\lambda_i)_{i=1}^{10}$ = (0.46, 0.45, 0.5, 0.48, 0.51, 0.4, 0.43, 0.42, 0.45,0.44) and $T = 10^5$.

Stats	BMLE	VIDS	KLUCB	TS	UCB	MOSS	BUCB	KG
Mean	**195.2**	215.9	339.1	221.3	1815.8	462.0	298.8	1460.3
SD	140.2	425.2	53.6	60.3	69.2	53.3	**45.9**	2035.8
Q10	140.9	43.5	264.9	159.6	1729.3	402.8	247.8	**26.7**
Q25	153.1	55.9	301.4	176.9	1776.9	428.6	263.5	**33.7**
Q50	166.2	70.5	335.7	211.3	1818.0	456.7	297.7	**58.3**
Q75	188.0	**94.1**	373.1	248.6	1863.7	480.2	326.5	3249.7
Q90	**225.8**	1037.1	408.4	296.9	1897.8	532.3	365.8	4955.2
Q95	**291.7**	1064.6	433.2	319.3	1934.1	563.1	383.3	4966.9

Table 2.10: Statistics of distribution of average final regret over 100 trials for the Exponential bandits with true values: $(1/\lambda_i)_{i=1}^{10}$ = (0.25, 0.28, 0.27, 0.3, 0.29, 0.22, 0.21,0.24, 0.23, 0.26) and $T = 10^5$.

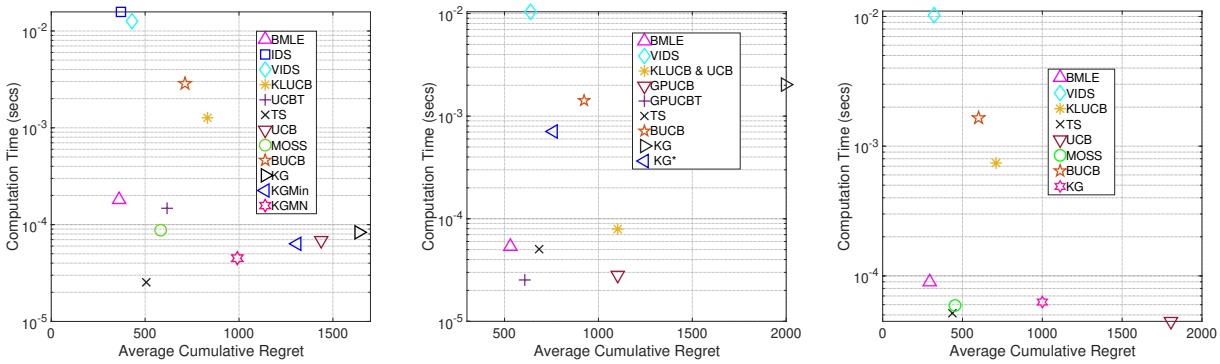
for an arm, except in Gaussian bandits, where a closed-form solution can be obtained. We observe from the figure that those baselines such as UCB, GPUCB, and KG that also enjoy closed-form index have similar computation time per decision as BMLE, i.e., comparable vertical position in



(a) (b) (c)

Figure 2.4: Comparison of computation time and regret for Bernoulli, Gaussian, and Exponential bandits over 100 trials with $T = 10^5$ and $\beta(t) = \sqrt{\log(t)}$: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} =$ (0.66, 0.67, 0.68, 0.69, 0.7, 0.61, 0.62, 0.63, 0.64, 0.65); (b) Gaussian bandits with $\sigma = 1$ and $(\mu_i)_{i=1}^{10} =$ (0.41, 0.52, 0.66, 0.43, 0.58, 0.65, 0.48, 0.67, 0.59, 0.63); (c) Exponential bandits with the rates for different arms $(\lambda_i)_{i=1}^{10} =$ (1/0.31, 1/0.1, 1/0.2, 1/0.32, 1/0.33, 1/0.29, 1/0.2, /0.3, 1/0.15, 1/0.08).

the figure. However, their regret performance is far worse than BMLE's, i.e., larger horizontal position in the Figure, thus worse efficiency than BMLE. We also observe that in terms of efficiency, the closest competitors to BMLE are TS, MOSS, and tuned version UCB (UCBT, GPUCBT). Compared to TS, BMLE follows the frequentist formulation, and thus its performance does not deteriorate like TS when an inappropriate prior is mistakenly chosen. Compared to MOSS, BMLE does not rely on the knowledge of T to compute its index. Compared to the tuned version UCB, the regret performance of BMLE enjoys stronger theoretical guarantees, as illustrated by the aforementioned propositions.



(a) (b) (c)

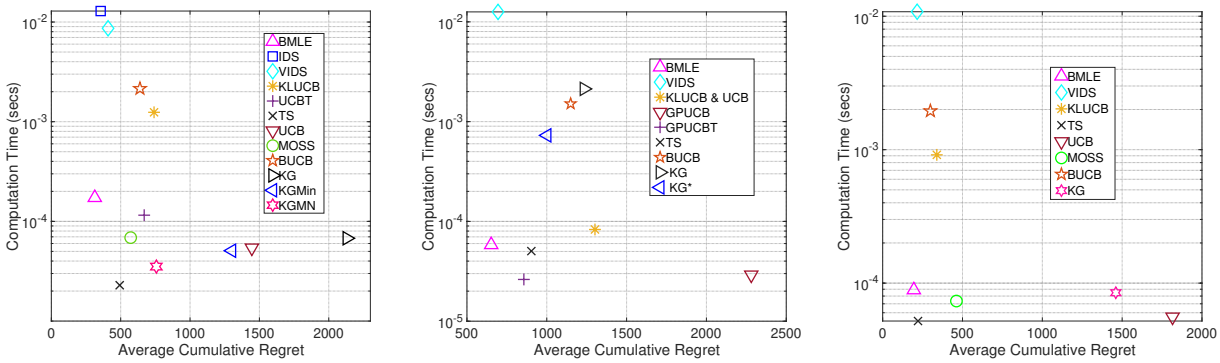
Figure 2.5: Comparison of computation time and regret for Bernoulli, Gaussian, and Exponential bandits over 100 trials with $T = 10^5$ and $\beta(t) = \sqrt{\log(t)}$: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} =$ (0.655, 0.6, 0.665, 0.67, 0.675, 0.68, 0.685, 0.69, 0.695, 0.7); (b) Gaussian bandits with $\sigma = 1$ and $(\mu_i)_{i=1}^{10} =$ (0.5, 0.75, 0.4, 0.6, 0.55, 0.76, 0.68, 0.41, 0.52, 0.67); (c) Exponential bandits with the rates for different arms $(\lambda_i)_{i=1}^{10} =$ (1/0.46, 1/0.45, 1/0.5, 1/0.48, 1/0.51, 1/0.4, 1/0.43, 1/0.42, 1/0.45, 1/0.44).

2.6.6 Scalability of the BMLE Algorithm

In this subsection, we compare the computation time per decision between BMLE and other methods when the number of arms increases. The computation times are measured on a Linux server with (i) an Intel Xeon E7 v4 server[2] operating at a maximal clock rate 3.60 GHz and (ii) a total of 528 GB memory. Throughout this section, we measure the average computation time per decision for each method over 100 simulation trials and a time horizon of 10000 for each trial. Tables 2.11-2.13 show the computation time per decision of different methods under varying numbers of arms. We observe that BMLE scales well for various reward distributions as the number of arms increases. The computation time per decision stays at a few 10^{-4} seconds even when the number of arms reaches 70. In contrast, the computation time per decision for VIDS and IDS can be as high as thousands of 10^{-4} seconds. The computation time for KLUCB, KG, and BUCB is often tens of times higher than BMLE. The only exception is the BUCB in Bernoulli bandits, where the quantile function is easier to be computed. The increased amount of time for IDS, VIDS,

[2]While there are 64 cores in the server, we force the program to run on just one core for a fair comparison.

KLUCB, and KG, is often much more than that for BMLE. It also deserves to be emphasized that, in Gaussian bandits, BMLE achieves the shortest computation time per decision when the number of arms is 30, 50, and 70. This is largely because, in Gaussian bandits, the computation of $\hat{C}_\alpha(t)$ is simple than in the other two types of bandits, as illustrated in Algorithm 3. We also observe that TS achieves the best performance in computation time in Bernoulli bandits and exponential bandits. The computation time of BMLE is often in the same order with that of TS and becomes closer and closer as the number of arms increases.



(a) (b) (c)

Figure 2.6: Comparison of computation time and regret for Bernoulli, Gaussian, and Exponential bandits over 100 trials with $T = 10^5$ and $\beta(t) = \sqrt{\log(t)}$: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} =$ (0.755, 0.76, 0.765, 0.77, 0.775, 0.78, 0.785, 0.79, 0.795, 0.8); (b) Gaussian bandits with $\sigma = 1$ and $(\mu_i)_{i=1}^{10} =$ (0.65, 0.35, 0.66, 0.4, 0.65, 0.64, 0.55, 0.4, 0.57, 0.54); (c) Exponential bandits with the rates for different arms $(\lambda_i)_{i=1}^{10} =$ (1/0.25, 1/0.28, 1/0.27, 1/0.3, 1/0.29, 1/0.22, 1/0.21, 1/0.24, 1/0.23, 1/0.26).

2.7 Possible Extensions

There are several promising directions to extend the proposed family of BMLE algorithms. One natural direction is to derive the index of BMLE algorithms and conduct regret analysis for in different contextual bandits. In this section, we mainly derive the BMLE index and quantify its performance for different context-free bandits. Considering its outstanding performance in

#Arms (Stats)	BMLE	IDS	VIDS	KLUCB	UCBT	TS	UCB	MOSS	BUCB	KG	KGMin	KGMN
10(Mean)	1.36	175	123	12.8	1.53	0.23	0.712	0.895	0.855	28.7	0.649	0.453
30(Mean)	3.61	1260	788	49.7	4.96	0.63	2.19	2.83	2.58	97.6	1.89	1.36
50(Mean)	4.58	3630	1930	80.3	7.85	0.63	3.42	4.40	4.11	159	2.95	2.14
70(Mean)	7.56	6660	3590	113	10.3	0.63	4.49	5.87	5.43	209	3.97	2.86
10(SE)	0.236	54.8	33.1	1.53	0.586	0.04	0.268	0.333	0.351	10.9	0.284	0.172
30(SE)	1.30	458	232	17.3	1.52	0.11	0.646	0.844	0.714	29.2	0.557	0.408
50(SE)	2.04	972	536	29.4	2.59	0.11	1.11	1.40	1.25	49.5	0.931	0.678
70(SE)	2.70	1330	883	36.6	3.63	0.11	1.53	2.00	1.76	69.3	1.34	0.962

Table 2.11: Average computation time per decision for Bernoulli bandits under different numbers of arms. All numbers are obtained over 100 trials with time horizon 10^4 and in 10^{-4} seconds.

#Arms (Stats)	BMLE	VIDS	KLUCB&UCB	GPUCB	GPUCBT	TS	BUCB	KG	KG*
10(Mean)	0.617	135	0.993	0.346	0.318	0.451	17.9	25.1	10.9
30(Mean)	1.07	1410	3.82	1.10	1.08	1.33	75.2	103	21.2
50(Mean)	1.49	3580	6.49	1.79	1.76	2.44	121	168	33.9
70(Mean)	1.95	6610	8.52	2.24	2.22	3.16	162	226	45.9
10(SE)	0.284	53.9	0.417	0.136	0.160	0.0425	6.98	9.37	2.77
30(SE)	0.484	409	1.28	0.370	0.370	0.321	26.2	35	5.61
50(SE)	0.686	866	2.14	0.563	0.563	0.562	42.1	56.1	9.77
70(SE)	0.871	1290	2.95	0.755	0.773	0.774	58.5	77.6	15.7

Table 2.12: Average computation time per decision for Gaussian bandits under different numbers of arms. All numbers are obtained over 100 trials with time horizon 10^4 and in 10^{-4} seconds.

#Arms (Stats)	BMLE	VIDS	KLUCB	TS	UCB	MOSS	BUCB	KG
10(Mean)	1.01	133	7.26	1.38	0.420	0.548	14.9	0.519
30(Mean)	1.93	1160	22.8	3.97	1.20	1.61	42.6	1.36
50(Mean)	2.97	3170	36.5	6.64	1.92	2.53	75.5	2.23
70(Mean)	3.79	6430	53.7	9.30	2.67	3.59	102	3.06
10(SE)	0.435	13.6	0.884	0.316	0.0980	0.112	1.55	0.101
30(SE)	0.890	187	2.79	0.777	0.263	0.340	5.02	0.265
50(SE)	1.24	447	5.47	1.20	0.397	0.498	10.2	0.456
70(SE)	1.56	788	6.92	1.96	0.531	0.688	12.3	0.605

Table 2.13: Average computation time per decision for Exponential bandits under varying numbers of arms. All numbers are obtained over 100 trials with time horizon 10^4 and in 10^{-4} seconds.

context-free scenarios, it will be interesting to examine how well the family of BMLE algorithms performs under contextual information. Another promising direction is in the areas of MDP and reinforcement learning. As mentioned in Section 2.2, the proposed BMLE algorithms strongly connect to the unknown MDP problem – a fundamental problem in adaptive control and reinforcement learning. As such, one possible direction is to extend the precise regret and computational analysis to general adaptive control of Markov chains, and the efficient exploration problem in reinforcement learning. There have been studies that successfully extend other bandit algorithms to solve the same problem [53, 54, 55, 56]. Another possible direction to extend the family of BMLE algorithm is in the areas of Bayesian Optimization, which in some sense can be viewed as a pure-exploration problem in "continuous" bandits. In view of many previous successful extensions of bandit algorithms to Bayesian Optimization [57, 58, 59], it will not be surprising if BMLE is also of interest in in Bayesian Optimization.

2.8 Summary

In this section, we propose BMLE – a novel family of bandit algorithms to overcome the limitation in efficiency in the best-performing algorithms for bandit learning. The proposed BMLE-family algorithms are formulated in a general way and are based on the Biased Maximum Likelihood Estimation method originally appearing in the adaptive control literature. Although a similar scheme appears in previous studies, it has never been considered in bandit setting with respect to the finer notion of regret. Here we design the reward-bias term to tackle the exploration and exploitation tradeoff for stochastic bandit problems and shown that it is a competitive method with performance often slightly better than other state-of-the-art baseline methods. Moreover, BMLE provides simple indices that provide a major computational advantage in terms of being very easy-to-compute for each arm. We prove that the derived BMLE indices achieve a logarithmic finite-time regret bound and hence attain order-optimality, for both exponential families and the cases beyond parametric distributions. Through extensive simulations, we demonstrate that the proposed algorithms achieve regret performance comparable to the best of several state-of-the-art baseline methods while being computationally efficient in comparison to other best-performing

methods. Unlike some other bandit learning algorithms that rely on a lot of intuitions to derive the index, the clear theoretical foundation and generality of the proposed family of BMLE algorithms potentially is expected to be extendable to several promising formulations.

3. LEARNING TO OPTIMIZE UNDER PRESENCE OF RENEGING RISK AND REWARD HETEROSCEDASTICITY[1]

3.1 Overview

In this section, we introduce HR-UCB – a novel learning algorithm for contextual bandit problems. The bandit models that are discussed in Section 2 are usually referred to as *context*-free bandits. One of the major limitations in modeling real-world problems is that they do not use any applicable "features" in determining the values of unknown reward parameters. This makes them not as competitive as other learning algorithms in modeling problems with big data. To overcome this limitation, researchers in the bandit community have proposed "contextual" bandits. Compared to context-free bandits, in contextual bandits, reward parameters are often assumed to depend on the contexts (features) of the bandits. For example, in a contextual bandit referred to as a "linear" bandit, the mean value of the reward distribution equals the dot product of the features and an unknown coefficient vector to be learned on the fly [26].

Sequential decision problems commonly arise in a large number of real-world applications. To name a few, in treatments to extend the life of people with terminal illnesses, doctors are required to make decisions on which treatments are to be used for patients periodically. In portfolio selection, fund managers need to decide which portfolios are recommended to their customers every time. In cloud computing services, the cloud platform has to determine the resources allocated to customers given specific requirements of their programs. Contextual bandits [27] have been extensively used to model such problems. In the modeling, available choices are referred to as "arms" and a decision is regarded as a "pull" of the corresponding arm. The decision is evaluated through rewards that depend on the outcome of the interaction.

In the aforementioned applications of contextual bandits, the phenomenon of participants disengage from future interactions has been commonly observed. Such behavior is referred to as

[1]Part of this section is reprinted with permission from "Stay With Me: Lifetime Maximization Through Heteroscedastic Linear Bandits With Reneging" in ICML 2019 [60] Copyright by the authors themselves Ping-Chun Hsieh*, Xi Liu*, Anirban Bhattacharya and P. R. Kumar. *: Equal contribution.

"churn", "unsubscribing" or "reneging" in the literature [61, 62]. For instance, patients may fail to survive the illness or are unable to undertake more treatments due to the deterioration of their physical condition [63]. In portfolio selection, fund managers earn money from customer enrollment in their service. The return from the portfolio selected may however turn out to be loss, occasioning the customer to lose trust in the manager and stoppage of using the service [64]. Similarly, in cloud computing services, the customer may feel that a resource was not well allocated and be dissatisfied with the throughput, and then switch to another service provider [65]. In other words, the participant [2] of the interaction often has a limited "lifetime" defined as the total number of interactions between the participant and a service provider until the customer reneges. The larger the lifetime, the "longer" the participant stays with the provider. Customer lifetime has been recognized as a critical metric to evaluate the success of many applications including all the aforementioned applications as well as e-commerce applications [66]. Moreover, as well known, the acquisition cost for a new customer is much higher than an existing customer [61]. Therefore, in such applications and services, a particularly vital goal is to maximize the lifetime of customers. Unfortunately, this reneging risk is rarely discussed in existing bandit solutions. Most existing bandit algorithms assume that the interaction process never ends. Their objective is only to maximize the accumulated rewards collected from endless interactions. As such, they are not directly applicable to the problem of customer reneging.

Another phenomenon that has been neglected in many contextual bandit formulations is the presence of "heteroscedasticity" in real-world applications, i.e., the variability of outcomes across the range of predictors. Many previous studies of the aforementioned applications have pointed out that the distribution of the outcome can be heteroscedastic. In medical treatment of patients, it has been found that the physical condition after treatment can be highly heteroscedastic [16, 17]. In portfolio selection [18, 19, 20], it is even more common that the return of investing in a selected portfolio is heteroscedastic. In cloud services, it has been repeatedly observed that throughput and responses of the servers can be highly heteroscedastic [67, 68, 69]. In the bandit

[2]For simplicity, in this section, we use the terms *participant*, *user*, *customer*, and *patients* interchangeably.

setting, this means that *both* the mean value and the variance of the outcome depend on the context. The "context" here represents both the decision and the customer. However, previous studies on contextual bandits have usually assumed that the underlying distribution involved in the problem is homoscedastic, i.e., its variance is independent of contexts. As such, they only need to estimate the true value of the mean. If the reneging risk is the chance that the outcome (e.g., patients' health condition, portfolio return, and throughput rate) is below the satisfaction level, accurately estimating it requires estimation of both mean and variance. Existing contextual bandit algorithms are therefore inapplicable under the two phenomena.

The line of MAB research that is most relevant to the problem is bandit models with risk management, e.g., variance minimization [70] and value-at-risk maximization [71, 72, 73]. However, the risks in those models concern the large fluctuation of collected rewards which have no impact on the lifetimes of bandits. This renders them inapplicable to our problem. Another category of related research is "conservative" bandits [74, 75], where a choice is only considered if it guarantees that the *overall* performances outperforms $1-\alpha$ of baselines. Unfortunately, our problem has a higher degree of granularity, i.e., to avoid reneging, *individual* performance (performance of each choice) needs to be above some satisfaction level. Moreover, none of them considers data heteroscedasticity. A more complete review and comparison are provided in Section 3.2.

To overcome these limitations, we propose a novel model of contextual bandits that addresses the challenges arising from both reneging risk as well as heteroscedasticity. We call the model "heteroscedastic linear bandits with reneging". To solve the proposed model, we develop a UCB-type policy, called Heteroscedastic Risk Upper Confidence Bounds (HR-UCB), that is proved to achieve a $O\big(\sqrt{T(\log(T))^3}\big)$ regret bound with high probability. We have successfully applied the proposed method to solve the lifetime maximization problem. We evaluate the performance of HR-UCB for the problem via comprehensive simulations. The simulation results demonstrate that our model has lower regret, and outperforms conventional UCB that ignores reneging, as well as more complex models such as Episodic Reinforcement Learning (ERL). The main contributions of this section are as follows:

- The reneging risk and reward heteroscedasticity commonly arise in many real-world applications of bandit learning but are ignored by most existing bandit models. We investigate the characteristics of the two phenomena and propose a novel bandit model that is able to overcome the limitation of bandit models in the presence of reneging risk and reward heteroscedasticity.

- To provide a solution for the proposed model, we develop a UCB-type policy, called HR-UCB, and establish theoretical guarantee for the proposed policy. We prove that the HR-UCB can achieve a $O\big(\sqrt{T(\log(T))^3}\big)$ regret bound with high probability.

- We evaluate the HR-UCB via comprehensive simulations. The simulation results demonstrate that the model outperforms conventional UCB that ignores reneging and more complex models such as Episodic Reinforcement Learning (ERL).

3.2 Related Work

There are mainly two lines of research related to our work. The first is about bandits with risk management. Reneging can be viewed as a type of risk that the decision-maker tries to avoid. The risk management in bandit problems has been studied in terms of variance and quantiles. In [70], mean-variance models to handle risk are studied, where the risk refers to the variability of collected rewards. The difference from conventional bandits is that the objective to be maximized is a linear combination of mean reward and variance. Subsequent studies [71, 72] propose a quantile (value at risk) to replace the mean-variance objective. While these studies investigate optimal policies under risk, the risks they handle are different from ours, in the sense that the risks usually relate to the variability of rewards and have no impact on the lifetime of bandits. Moreover, their approaches to handle the risk are based on more straightforward statistics, while, in our problem, the reneging risk is relatively complex, i.e., it comes from the probability that the outcome of following a suggestion is below a satisfactory level. Therefore, their models cannot be used to solve our problem.

Second, in contrast to those works, *conservative bandits* [74, 75] control the risk by requiring that the accumulated rewards while learning the optimal policy be above those of baselines.

Similarly, in [76], each arm is associated with some risk; safety is guaranteed by requiring the accumulated risk to be below a given budget. Unfortunately, our problem has a higher degree of granularity. The participants in our problem are more sensitive to bad suggestions. A single bad decision may cause reneging and brings the interactions to an end, e.g., one bad treatment can result in a patient's death.. Moreover, their models assume homoscedasticity, while we allow the variance to depend on the context.

The "satisfaction level" in our model has the flavor of thresholding bandits. Different from us, the thresholds in the existing literature are mostly used to model reward generation. For instance, in [77], an action induces a unit payoff if the sampled outcome exceeds a threshold. In [78], no rewards can be collected until the total number of successes exceeds the threshold.

In terms of the problem in this section, the most relevant one that has previously been studied is in [79]. Compared to it, ours has three salient differences. First, it has a very different setting for modeling reneging: each decision is represented by a real number; reneging happens when the pulled arm falls below a threshold. As a comparison, we represent each decision by a high-dimensional context vector; reneging happens if the outcome of following a suggestion is not satisfactory. Second, it couples the reneging with the reward generation. The "rewards" in our model can be regarded as the lifetime while the reneging is separately captured by the outcome distribution. Third, it fails to take into account the data heteroscedasticity in the aforementioned applications.

In terms of bandits under heteroscedasticity, to the best of our knowledge, only one very recent paper [80] discusses it. Compared to it, ours has two salient differences. First, we address heteroscedasticity under the presence of reneging. The presence of reneging makes the learning problem more challenging as the learner has to always be prepared that plans for the future may not be carried out. Second, the solution in [80] is based on information directed sampling. In contrast, in this section, we present a heteroscedastic UCB policy that is efficient, easier to implement, and can achieve sub-linear regret. The reneging problem can also be approximated by an infinite-horizon Episodic Reinforcement Learning (ERL) problem [81, 82]. Compared to it, our

solution has two distinct features: (a) the reneging behavior and heteroscedasticity are explicitly addressed, (b) the context information is leveraged in learning policy design.

3.3 Problem Formulation

In this section, we describe the formulation of the heteroscedastic linear bandits with reneging. To incorporate reneging behavior into the bandit model, we model the problem in the following stylized manner: The users arrive at the decision-maker one after another and are indexed by $t = 1, 2, \cdots$. For each user t, the decision-maker interacts with the user in discrete *rounds* by selecting one action in each round sequentially until the user t reneges on interacting with the decision-maker. Let s_t denote the total number of rounds experienced by the user t. Note that s_t is a stopping time, which depends on the reneging mechanism that will be described shortly. Since the decision-maker interacts with one user at a time, all the actions and the corresponding outcomes regarding user t are determined and observed, before the next user, $t + 1$ arrives.

Let $\mathcal{A}$ be the set of available actions of the decision-maker. Upon the arrival of each user t, the decision-maker observes a set of *contexts* $\mathcal{X}_t = \{x_{t,a}\}_{a \in \mathcal{A}}$, where each context $x_{t,a} \in \mathcal{X}_t$ summarizes the pair-wise relationship[3] between the user t and the action a. Without loss of generality, we assume that for any user t and any action a, we have $\|x_{t,a}\|_2 \leq 1$, where $\|\cdot\|_2$ denotes the ℓ_2-norm. After observing the contexts, the decision-maker selects an action $a \in \mathcal{A}$ and observes a random outcome $r_{t,a}$. We assume that the outcomes $r_{t,a}$ are conditionally independent random variables given the contexts, and are drawn from an outcome distribution that satisfies:

$$r_{t,a} := \theta_*^\top x_{t,a} + \varepsilon(x_{t,a}) \tag{3.1}$$

$$\varepsilon(x_{t,a}) \sim \mathcal{N}\big(0, \sigma^2(x_{t,a})\big) \tag{3.2}$$

$$\sigma^2(x_{t,a}) := f(\phi_*^\top x_{t,a}), \tag{3.3}$$

where $\mathcal{N}(0, \sigma^2)$ denotes the Gaussian distribution with zero mean and variance σ^2, and $\theta_*, \phi_* \in \mathbb{R}^d$

[3]For example, in recommender systems, one way to construct such a pair-wise context is to concatenate the feature vectors of each individual user and each individual action.

are unknown, but known to have the norm bounds as $||\theta_*||_2 \leq 1$ and $||\phi_*||_2 \leq L$. Although, for simplicity of discussion, we focus here on Gaussian noise, all of our analysis can be extended to sub-Gaussian outcome distributions of the form $\psi_\sigma(x) = (1/\sigma)\psi((x-\mu)/\sigma)$, where ψ is a known sub-Gaussian density with unknown parameters μ, σ. This family includes truncated distributions and mixtures, thus allowing multi-modality and skewness. The parameter vectors $\theta_* \in \mathbb{R}^d$ and $\phi_* \in \mathbb{R}^d$ will be learned by the decision-maker during interactions with the users. The function $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is assumed to be a known linear function with a finite positive slope M_f such that $f(z) \geq 0$, for all $z \in [-L, L]$. One example that satisfies the above conditions is $f(z) = z + L$. Note that the mean and variance of the outcome distribution satisfy

$$\mathbb{E}[r_{t,a}|x_{t,a}] := \theta_*^\top x_{t,a}, \tag{3.4}$$

$$\mathbb{V}[r_{t,a}|x_{t,a}] := f(\phi_*^\top x_{t,a}). \tag{3.5}$$

Since $\phi_*^\top x_{t,a}$ is bounded over all possible ϕ_* and $x_{t,a}$, we know that $f(\phi_*^\top x_{t,a})$ is also bounded, i.e. $f(\phi_*^\top x_{t,a}) \in [\sigma_{\min}^2, \sigma_{\max}^2]$ for some $\sigma_{\min}, \sigma_{\max} > 0$, for all ϕ_* and $x_{t,a}$ defined above. This also implies that $\varepsilon(x_{t,a})$ is $\sigma_{\max}^2$-sub-Gaussian, for all $x_{t,a}$.

3.3.1 Model of Reneging Behavior

We modeled reneging behavior based on two observations. First, in all the applications mentioned in Section 3.1, the decision-maker is usually able to observe the outcome of following the suggestion, e.g., the physical condition of the patient after the treatment, the money earned from purchasing the suggested portfolio, and the throughput rate of running the programs. Second, we observe that the participants in these applications are willing to reveal their satisfaction level with respect to the outcome of the suggestion. For instance, patients will let doctors know their expectations for the treatment in physician visits. Customers are willing to inform fund managers how much money they can afford to lose. Cloud users share with the service providers their requirements of throughput performance. We suppose that the outcome of following the suggestion is a random variable drawn from an unknown distribution that may vary under different contexts. If

the outcome of the random variable falls below the satisfaction level, the customer quits all future interactions, i.e., "reneges"; otherwise, the customer stays. The reneging risk is, therefore, the chance that the outcome drawn from an unknown distribution falls below some customized threshold. Thus, learning the unknown outcome distribution plays a critical role in optimal decision making.

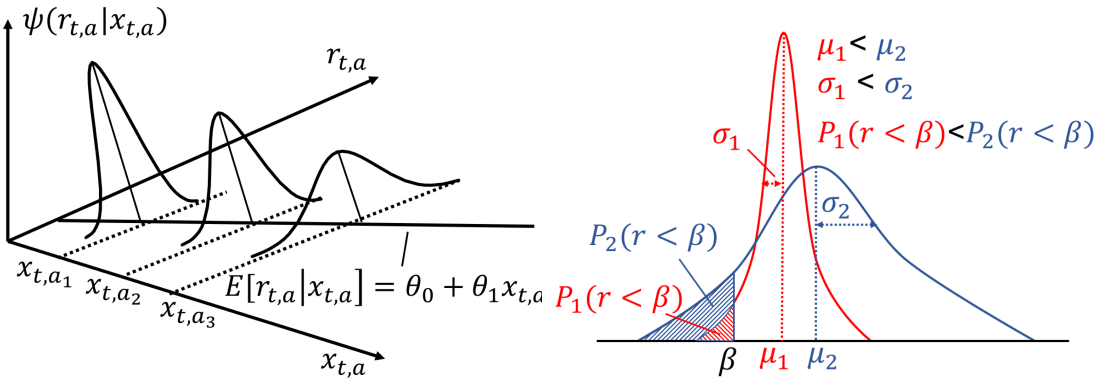
The minimal expectation of a user is characterized by its *satisfaction level*. Let $\beta_t \in \mathbb{R}$ denote the "satisfaction level" of user t. We assume that satisfaction levels of users, like the pair-wise contexts, are available before interacting with them. Denote by $r_t^{(i)}$ the observed outcome at round i of user t. When $r_t^{(i)}$ falls below β_t, reneging occurs and the user drops out from any future interaction. Supposing that at round i, action a is selected for user t, the risk that reneging occurs is

$$\mathbb{P}(r_t^{(i)} < \beta_t | x_{t,a}) = \Phi\Big(\frac{\beta_t - \theta_*^\top x_{t,a}}{\sqrt{f(\phi_*^\top x_{t,a})}}\Big), \tag{3.6}$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) for $\mathcal{N}(0,1)$. Without loss of generality, we also assume that β_t is lower bounded by $-B$ for some $B > 0$. Recall that s_t denotes the number of rounds experienced by user t. Given the reneging behavior as modeled above, s_t is the stopping time that represents the first time that the outcome $r_t^{(i)}$ is below the satisfaction level β_t, i.e. $s_t := \min\{i : r_t^{(i)} < \beta_t\}$.

3.3.2 Model of Heteroscedasticity

Illustrative examples of heteroscedasticity and reneging risk are shown in Figure 3.1. In Figure 3.1(a), the variance of the outcome distribution gradually increases as the value of the one-dimensional context $x_{t,a}$ increases. Figure 3.1(b) shows the outcome distributions of the two actions for a user. Specifically, the outcome distribution $\mathbb{P}_1$ has mean μ_1 and variance σ_1^2, and mean μ_2 and variance σ_2^2 for $\mathbb{P}_2$. As the two distributions correspond to the same user (but for different actions), they face the same satisfaction level β. In this example, the reneging risk $\mathbb{P}_2(r < \beta)$ (the blue shaded area) is higher than $\mathbb{P}_1(r < \beta)$ (the red shaded area).



(a) Example of heteroscedasticity

(b) Example of reneging risk

Figure 3.1: Illustrative examples of heteroscedasticity and reneging risk in the presence of heteroscedasticity. ($\psi(\cdot)$ is the probability density function.)

A policy $\pi \in \Pi$ is a rule for selecting an action at each round for a user based on the preceding interactions with that user and other users, where Π denotes the set of all admissible policies. Let $\pi_t = \{x_{t,1}, x_{t,2}, \cdots\}$ denote the sequence of contexts that correspond to the actions for user t under policy π. To solve the *lifetime maximization* problem, let $\overline{R}_t^{\pi}$ denote the expected lifetime of user t under the action sequence π_t. Then the total expected lifetime of T users can be represented by $\mathcal{R}^{\pi}(T) = \sum_{t=1}^{T} \overline{R}_t^{\pi_t}$. Define π^* as the optimal policy in terms of total expected lifetime among admissible policies, i.e. $\pi^* = \text{argmax}_{\pi \in \Pi} \mathcal{R}^{\pi}(T)$. The *pseudo regret* of the heteroscedastic linear bandits with reneging for a policy π is

$$\text{Regret}_T := \mathcal{R}^{\pi^*}(T) - \mathcal{R}^{\pi}(T). \tag{3.7}$$

The objective of the decision-maker is to learn a policy that achieves as minimal a regret as possible.

3.4 The HR-UCB Algorithm

3.4.1 Oracle Policy

Before we propose our policy, let us first consider how an oracle policy with full knowledge of θ_* and ϕ_* will make a decision. Consider T users that arrive sequentially. Let $\pi_t^{\text{oracle}} = \{x_{t,1}^*, x_{t,2}^*, \cdots\}$ be the sequence of contexts that correspond to the actions for the user t under an oracle policy π^{oracle}. The oracle policy $\pi^{\text{oracle}} = \{\pi_t^{\text{oracle}}\}$ is constructed by choosing

$$\pi_t^{\text{oracle}} = \arg\max_{\widetilde{x}_t = \{\widetilde{x}_{t,1}, \widetilde{x}_{t,2} \cdots\}} R_t^{\widetilde{x}_t}, \tag{3.8}$$

for each t. Due to the construction in (3.8), we know that π^{oracle} achieves the largest possible expected lifetime for each user t, and is hence optimal in terms of pseudo-regret defined in Section 3.3. By using an one-step optimality argument, it is easy to verify that π^{oracle} is a fixed policy for each user t, i.e. $x_{t,i} = x_{t,j}$, for all $i, j \geq 1$. Let $\overline{R}_t^*$ denote the expected lifetime of user t under π^{oracle}. As such, the optimal reward an oracle policy can receive from user t is

$$\overline{R}_t^* = \left(\Phi\Big(\frac{\beta_t - \theta_*^\top x_t^*}{\sqrt{f(\phi_*^\top x_t^*)}}\Big)\right)^{-1}. \tag{3.9}$$

This consideration to the oracle policy inspires us to propose HR-UCB replacing θ_* and ϕ_* with their estimation. The challenges will be how to handle the heteroscedasticity and different lifetimes of users in the estimation, as well as quantify the regret performance with the replacement.

3.4.2 Estimators for θ_* and ϕ_*

Consider a general regression problem with heteroscedasticity. Let $\{(x_i, r_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^n$ be a sequence of n pairs of context and outcome that are realized by a user's actions. Recall from (3.1)-(3.3) that $r_i = \theta_*^\top x_i + \varepsilon(x_i)$ and $\varepsilon(x_i) \sim \mathcal{N}\big(0, f(\phi_*^\top x_i)\big)$ with unknown parameters θ_* and ϕ_*. Note that, given the contexts $\{x_i\}_{i=1}^n$, $\varepsilon(x_1), \cdots, \varepsilon(x_n)$ are mutually independent. Let $r = (r_1, \cdots, r_n)^\top$ and $\varepsilon = (\varepsilon(x_1), \cdots, \varepsilon(x_n))$ be the row vectors of the n outcome realizations and the deviations from the mean, respectively. Let $\boldsymbol{X}_n$ be an $n \times d$ matrix in which the i-th row

is $x_i^\top$, for all $1 \leq i \leq n$. We use $\widehat{\theta}_n, \widehat{\phi}_n \in \mathbb{R}^d$ to denote the estimators of θ_* and ϕ_* based on the observations $\{(x_i, r_i)\}_{i=1}^n$, respectively. Moreover, define the estimated residual with respect to $\widehat{\theta}_n$ as $\widehat{\varepsilon}(x_i) = r_i - \widehat{\theta}_n^\top x_i$. Let $\widehat{\varepsilon} = (\widehat{\varepsilon}(x_1), \cdots, \widehat{\varepsilon}(x_n))^\top$. Let $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix, and let $z_1 \circ z_2$ denote the Hadamard product of any two vectors z_1, z_2. We consider the *generalized least squares estimators* (GLSE) [83] as

$$\widehat{\theta}_n = \big(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}_d\big)^{-1} \boldsymbol{X}_n^\top r, \tag{3.10}$$

$$\widehat{\phi}_n = \big(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}_d\big)^{-1} \boldsymbol{X}_n^\top f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}), \tag{3.11}$$

where $\lambda > 0$ is some regularization parameter and $f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) = (f^{-1}(\widehat{\varepsilon}(x_1)^2), \cdots, f^{-1}(\widehat{\varepsilon}(x_n)^2))^\top$ is the pre-image of the vector $\widehat{\varepsilon} \circ \widehat{\varepsilon}$. Note that in (3.10), $\widehat{\theta}_n$ is the conventional ridge regression estimator. On the other hand, to obtain an estimator $\widehat{\phi}_n$, (3.11) still follows the ridge regression approach, but with two additional steps: (i) derive the estimated residual $\widehat{\varepsilon}$ based on $\widehat{\theta}_n$, and (ii) apply the map $f^{-1}(\cdot)$ on the square of $\widehat{\varepsilon}$.

3.4.3 Pseudo Code of the HR-UCB Algorithm

Define a $d \times d$ matrix $\boldsymbol{V}_n$ as

$$\boldsymbol{V}_n = \big(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}_d\big). \tag{3.12}$$

For all $n \in \mathbb{N}$, define

$$\alpha_n^{(1)}(\delta) = \sigma_{\max}^2 \sqrt{d \log\Big(\frac{n+\lambda}{\delta\lambda}\Big)} + \lambda^{1/2}, \tag{3.13}$$

and $\|x\|_{\boldsymbol{V}_n} = \sqrt{x^\top \boldsymbol{V}_n x}$ is the induced vector norm of vector x with respect to $\boldsymbol{V}_n$. Define

$$\alpha^{(2)}(\delta) = \sqrt{2d(\sigma_{\max}^2)^2 \Big(\Big(\frac{1}{C_2} \ln(\frac{C_1}{\delta})\Big)^2 + 1\Big)}, \tag{3.14}$$

$$\alpha^{(3)}(\delta) = \sqrt{2d\sigma_{\max}^2 \ln\big(\frac{d}{\delta}\big)}, \tag{3.15}$$

where C_1 and C_2 are some universal constants. For all $n \in \mathbb{N}$, define

$$\rho_n(\delta) = \frac{1}{M_f}\Big\{\alpha_n^{(1)}(\frac{\delta}{3})\big(\alpha_n^{(1)}(\frac{\delta}{3}) + 2\alpha^{(3)}(\frac{\delta}{3})\big) \tag{3.16}$$

$$+ \alpha^{(2)}(\frac{\delta}{3})\Big\} + L^2\lambda^{1/2}. \tag{3.17}$$

For any given $\beta \in [-B, \infty)$, define the function $h_\beta : [-1, 1] \times [\sigma^2_{\min}, \sigma^2_{\max}] \to \mathbb{R}$ as

$$h_\beta(u, v) = \left(\Phi\Big(\frac{\beta - u}{\sqrt{f(v)}}\Big)\right)^{-1}. \tag{3.18}$$

Note that for any given $x \in \mathcal{X}$, $h_\beta({\theta_*}^\top x, {\phi_*}^\top x)$ equals the expected lifetime of a single user with threshold β if a fixed action with context x is chosen under parameters θ_*, ϕ_*. Note that in our bandit model, the number of rounds of each user is a stopping time and can be arbitrarily large. To address this, we propose to actively maintain a *regression sample set* $\mathcal{S}$ through a function $\Gamma(t)$. Specifically, we let the size of $\mathcal{S}$ grow at a proper rate regulated by $\Gamma(t)$. One example is to choose $\Gamma(t) = Kt$ for some constant $K \geq 1$. Since each user will play for at least one round, we know $|\mathcal{S}|$ is at least t after interacting with t users. We use $\mathcal{S}(t)$ to denote the regression sample set right after the departure of user t. Moreover, let $\boldsymbol{X}_t$ be the matrix in which the rows are composed by the contexts of all the elements in $\mathcal{S}(t)$. Similar to (3.12), we define $\boldsymbol{V}_t = \boldsymbol{X}_t^\top \boldsymbol{X}_t + \lambda \boldsymbol{I}_d$, for all $t \geq 1$. To simplify notation, we also define

$$\xi_t(\delta) := C_3\alpha^{(1)}_{|\mathcal{S}(t)|}(\delta) + C_4\rho_{|\mathcal{S}(t)|}(\delta/|\mathcal{S}(t)|^2). \tag{3.19}$$

Now we are ready to define the index of HR-UCB for any $x \in \mathcal{X}$:

$$Q^{\text{HR}}_{t+1}(x) := h_{\beta_{t+1}}\big(\widehat{\theta}_t^{\,\top} x, \widehat{\phi}_t^{\,\top} x\big) + \xi_t(\delta) \cdot \|x\|_{\boldsymbol{V}_t^{-1}}. \tag{3.20}$$

Note that $Q^{\text{HR}}_t(x)$ is indeed an upper confidence bound as will be illustrated in Section 3.5. Now, we formally introduce the HR-UCB algorithm in Algorithm 4.

Algorithm 4 The HR-UCB Algorithm

$\mathcal{S} \leftarrow \varnothing$, action set $\mathcal{A}$, function $\Gamma(t)$, and T
for each user $t = 1, 2, \cdots, T$ **do**
  observe $x_{t,a}$ for all $a \in \mathcal{A}$ and reset $i \leftarrow 1$
  while user t stays **do**
    $\pi_t^{(i)} = \arg\max_{x_{t,a} \in \mathcal{X}_t} Q_t^{\text{HR}}(x_{t,a})$ (ties are broken arbitrarily)
    apply the action $\pi_t^{(i)}$ and observe the outcome $r_t^{(i)}$ and if the reneging event occurs
    if $|\mathcal{S}| < \Gamma(t)$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(x_{t,\pi_t^{(i)}}, r_t^{(i)})\}$
    end if
    $i \leftarrow i + 1$
  end while
  update $\widehat{\theta}_t$ and $\widehat{\phi}_t$ by (3.10)-(3.11) based on $\mathcal{S}$
end for

As illustrated in Algorithm 4, for each user t, HR-UCB observes the contexts of all available actions, and then chooses an action based on the indices Q_t^{HR} that depend on $\widehat{\theta}_t$ and $\widehat{\phi}_t$. To derive these estimators by (3.10) and (3.11), HR-UCB actively maintains a sample set $\mathcal{S}$, whose size is regulated by a function $\Gamma(t)$. After applying an action, HR-UCB observes the corresponding outcome and the reneging event, if any. The current context-outcome pair will be added to $\mathcal{S}$ only if the size of $\mathcal{S}$ is less than $\Gamma(t)$. Based on the regression sample set $\mathcal{S}$, HR-UCB updates $\widehat{\theta}_t$ and $\widehat{\phi}_t$ right after the departure of each user. By using a one-step optimality argument, it is easy to verify that the optimal policy is a fixed policy for each user t, i.e. $x_{t,i} = x_{t,j}$, for all $i, j \geq 1$. This indicates that the exploration guaranteeing sublinear regret under heteroscedasticity is mainly over users. The knowledge transfer across users is given more importance than learning for a single user, because, compared to the population of potential users, a user's lifetime is mostly short. The concern of exploration is handled by encoding the confidence bound in Q_t^{HR} so that later users with similar contexts are treated differently.

3.5 Regret Analysis of the HR-UCB Algorithm

In this section, we provide regret analysis for HR-UCB.

3.5.1 Confidence Set of the Estimator for θ_*

First, let us see why $Q_t^{\mathrm{HR}}(x)$ is indeed an upper confidence bound and how tight it is. A confidence set for $\widehat{\theta}_*$ was introduced in [27]. For convenience, we restate these elegant results in the following lemma.

Lemma 5. (Theorem 2 in [27]) *For all $n \in \mathbb{N}$ and any $\delta > 0$, we have*

$$\mathbb{P}\Big\{\Big\|\widehat{\theta}_n - \theta_*\Big\|_{\boldsymbol{V}_n} \leq \alpha_n^{(1)}(\delta), \forall n \in \mathbb{N}\Big\} \geq 1 - \delta, \tag{3.21}$$

3.5.2 Confidence Set of the Estimator for ϕ_*

Next, we derive the confidence set for the estimator of ϕ_*. The following is the main theorem on the confidence set for $\widehat{\phi}_n$.

Theorem 1. *For all $n \in \mathbb{N}$ and for any $\delta > 0$, with probability at least $1 - 2\delta$, we have*

$$\Big\|\widehat{\phi}_n - \phi_*\Big\|_{\boldsymbol{V}_n} \leq \rho_n(\frac{\delta}{n^2}) = O\Big(\log(\frac{1}{\delta}) + \log n\Big), \forall n \in \mathbb{N}. \tag{3.22}$$

Remark 5. As the estimator $\hat{\phi}_n$ depends on the residual term $\hat{\varepsilon}$, which involves the estimator $\hat{\theta}_n$, it is expected that the convergence speed of $\hat{\phi}_n$ would be no larger than that of $\hat{\theta}_n$. Based on Theorem 1 along with Lemma 5, we know that under GLSE, $\hat{\phi}_n$ converges to the true value at a slightly slower rate than $\hat{\theta}_n$.

To demonstrate the main idea behind Theorem 1, we highlight the proof in the following Lemma 6-9. We start by taking the inner products of an arbitrary vector x with $\widehat{\phi}_n$ and ϕ_* to quantify the difference between $\widehat{\phi}_t$ and ϕ_*.

Lemma 6. *For any $x \in \mathbb{R}^d$, we have*

$$|x^\top \widehat{\phi}_n - x^\top \widehat{\phi}_*| \le \|x\|_{\boldsymbol{V}_n{}^{-1}} \big\{ \lambda \|\phi_*\|_{\boldsymbol{V}_n^{-1}} \tag{3.23}$$

$$+ \big\| \boldsymbol{X}_n^\top \big(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n \phi_* \big) \big\|_{\boldsymbol{V}_n{}^{-1}} \tag{3.24}$$

$$+ \frac{2}{M_f} \Big\| \boldsymbol{X}_n^\top \big(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \big) \Big\|_{\boldsymbol{V}_n{}^{-1}} \tag{3.25}$$

$$+ \frac{1}{M_f} \Big\| \boldsymbol{X}_n^\top \big(\boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \big) \Big\|_{\boldsymbol{V}_n{}^{-1}} \big\}. \tag{3.26}$$

Proof. The proof is provided in Appendix B.1. □

Based on Lemma 6, we provide upper bounds for the three terms in (3.24)-(3.26) separately as follows.

Lemma 7. *For any $n \in \mathbb{N}$, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$M_f \big\| \boldsymbol{X}_n^\top \big(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n \phi_* \big) \big\|_{\boldsymbol{V}_n{}^{-1}} \le \alpha^{(2)}(\delta). \tag{3.27}$$

Proof. We highlight the main idea of the proof. Recall that $\varepsilon(x_i) \sim \mathcal{N}(0, \phi_*^\top x_i)$. Therefore, $\varepsilon(x_i)^2$ is a χ_1^2-distribution with a scaling of $f(\phi_*^\top x_i)$. Hence, each element in $(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n \phi_*)$ has zero mean. Moreover, we observe that $\big\| \boldsymbol{X}_n^\top \big(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n \phi_* \big) \big\|_{\boldsymbol{V}_n{}^{-1}}$ is quadratic. Since the χ_1^2-distribution is sub-exponential, we utilize a proper tail inequality for quadratic forms of sub-exponential distributions to derive an upper bound. The complete proof is provided in Appendix B.2. □

Then, we derive an upper bound for (3.25).

Lemma 8. *For any $n \in \mathbb{N}$, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\Big\| \boldsymbol{X}_n^\top \big(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \big) \Big\|_{\boldsymbol{V}_n{}^{-1}} \le \alpha_n^{(1)}(\delta) \cdot \alpha^{(3)}(\delta). \tag{3.28}$$

Proof. The main challenge is that (3.28) involves the product of the residual ε and the estimation

error $\theta_* - \widehat{\theta}_n$. Through some manipulation, we can decouple ε from $\left\|\boldsymbol{X}_n^\top\big(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\big)\right\|_{\boldsymbol{V}_n{}^{-1}}$ and apply a proper tail inequality for quadratic forms of sub-Gaussian distributions. The complete proof is provided in Appendix B.3. □

Next, we provide an upper bound for (3.26).

Lemma 9. *For any $n \in \mathbb{N}$, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\left\|\boldsymbol{X}_n^\top\big(\boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\big)\right\|_{\boldsymbol{V}_n{}^{-1}} \leq (\alpha_n^{(1)}(\delta))^2. \tag{3.29}$$

Proof. Since (3.29) does not involve ε, we can simply reuse the results in Lemma 5 through some manipulation of (3.29). The complete proof is provided in Appendix B.4. □

Now, we are ready to prove Theorem 1.

Proof of Theorem 1. We use $\lambda_{\min}(\cdot)$ to denote the smallest eigenvalue of a square symmetric matrix. Recall that $\boldsymbol{V}_n = \lambda \boldsymbol{I}_d + \boldsymbol{X}_n^\top \boldsymbol{X}_n$ is positive definite for all $\lambda > 0$. We have

$$\|\phi_*\|^2_{\boldsymbol{V}_n{}^{-1}} \leq \|\phi_*\|_2^2 / \lambda_{\min}(\boldsymbol{V}_n) \leq \|\phi_*\|_2^2 / \lambda \leq L^2/\lambda. \tag{3.30}$$

By (3.30) and Lemmas 6-9, we know that for a given n and a given $\delta_n > 0$, with probability at least $1 - \delta_n$, we have

$$|x^\top \widehat{\phi}_n - x^\top \widehat{\phi}_*| \leq \|x\|_{\boldsymbol{V}_n{}^{-1}} \cdot \rho_n(\delta_n). \tag{3.31}$$

Note that (3.31) holds for any $x \in \mathbb{R}^d$. By substituting $x = \boldsymbol{V}_n(\widehat{\phi}_n - \phi_*)$ into (3.31), we have

$$\left\|\widehat{\phi}_n - \phi_*\right\|^2_{\boldsymbol{V}_n} \leq \left\|\boldsymbol{V}_n(\widehat{\phi}_n - \phi_*)\right\|_{\boldsymbol{V}_n{}^{-1}} \cdot \rho_n(\delta_n). \tag{3.32}$$

Since $\left\|\boldsymbol{V}_n(\widehat{\phi}_n - \phi_*)\right\|_{\boldsymbol{V}_n{}^{-1}} = \left\|\widehat{\phi}_n - \phi_*\right\|_{\boldsymbol{V}_n}$, we know for a given n and $\delta_n > 0$, with probability at

least $1-\delta_n$,

$$\left\|\widehat{\phi}_n - \phi_*\right\|_{\boldsymbol{V}_n} \leq \rho_n(\delta_n). \tag{3.33}$$

Finally, to obtain a uniform bound, we simply choose $\delta_n = \delta/(n^2)$ and apply the union bound to (3.33) over all $n \in \mathbb{N}$. Note that $\sum_{n=1}^{\infty} \delta_n = \sum_{n=1}^{\infty} \delta/n^2 = \frac{\pi^2}{6}\delta < 2\delta$. Therefore, with probability at least $1-2\delta$, for all $n \in \mathbb{N}$, $\left\|\widehat{\phi}_n - \phi_*\right\|_{\boldsymbol{V}_n} \leq \rho_n\big(\frac{\delta}{n^2}\big)$. □

3.5.3 Regret Proofs for the HR-UCB Algorithm

First, we show that $h_\beta(\cdot,\cdot)$ has the following nice property.

Theorem 2. *Let $\boldsymbol{M}$ be a $d \times d$ invertible matrix. For any $\theta_1, \theta_2 \in \mathbb{R}^d$ with $\|\theta_1\| \leq 1$, $\|\theta_2\| \leq 1$, for any $\phi_1, \phi_2 \in \mathbb{R}^d$ with $\|\phi_1\| \leq L$, $\|\phi_2\| \leq L$, for any $\beta \in [-B, \infty)$, $\forall x \in \mathcal{X}$,*

$$h_\beta\big(\theta_2^\top x, \phi_2^\top x\big) - h_\beta\big(\theta_1^\top x, \phi_1^\top x\big) \leq \tag{3.34}$$

$$\Big(C_3 \left\|\theta_2 - \theta_1\right\|_{\boldsymbol{M}} + C_4 \left\|\phi_2 - \phi_1\right\|_{\boldsymbol{M}}\Big) \cdot \|x\|_{\boldsymbol{M}^{-1}}\,, \tag{3.35}$$

where C_3 and C_4 are some finite positive constants that are independent of $\theta_1, \theta_2, \phi_1, \phi_2$, and β.

Proof. The main idea is to apply first-order approximation under Lipschitz continuity of $h_\beta(\cdot,\cdot)$. The detailed proof is provided in Appendix B.5. □

Then, we show that $Q_t^{\text{HR}}(x)$ is indeed an upper confidence bound.

Lemma 10. *If the confidence set conditions (3.21) and (3.22) are satisfied, then for any $x \in \mathcal{X}$,*

$$0 \leq Q_{t+1}^{\text{HR}}(x) - h_{\beta_{t+1}}\big(\theta_*^\top x, \phi_*^\top x\big) \leq 2\xi_t(\delta)\,\|x\|_{\boldsymbol{V}_t^{-1}}\,.$$

Proof. The proof is provided in Appendix B.6. □

Now, we formally provide regret analysis for the HR-UCB Algorithm.

Theorem 3. *Under HR-UCB, with probability at least* $1-\delta$*, the pseudo regret is upper bounded as*

$$\text{Regret}_T \leq \sqrt{8\xi_T^2(\frac{\delta}{3})T \cdot d\log\left(\frac{T+\lambda d}{\lambda d}\right)} \tag{3.36}$$

$$= O\left(\sqrt{T\log\Gamma(T)\cdot\left(\log\big(\Gamma(T)\big)+\log(\frac{1}{\delta})\right)^2}\right). \tag{3.37}$$

By choosing $\Gamma(T)=KT$ *with a constant* $K>0$*, we have*

$$\text{Regret}_T = O\left(\sqrt{T\log T\cdot\left(\log T+\log(\frac{1}{\delta})\right)^2}\right). \tag{3.38}$$

Proof. The proof is provided in Appendix B.7. □

Theorem 3 presents a high-probability regret bound. To derive an expected regret bound, we can set $\delta = 1/T$ in (3.37) and get $O(\sqrt{T(\log T)^3})$. Also note that the upper bound (3.36) depends on $\sigma_{\max}$ only through the pre-constant of ξ_T.

Remark 6. A policy that always assumes $\sigma_{\max}$ as variance tends to choose the action with the largest mean reward since it implies a smaller reneging probability. As a result, such type of policy incurs linear regret. This will be further demonstrated via simulations in Section 2.6.

Remark 7. The regret proof still goes through for sub-Gaussian noise by (a) reusing the same sub-exponential concentration inequality in Lemma B.1 since the square of a sub-Gaussian distribution is sub-exponential, (b) replacing the Gaussian concentration inequality in Lemma B.3 with a sub-Gaussian one, and (c) deriving ranges of the first two derivatives of sub-Gaussian CDF.

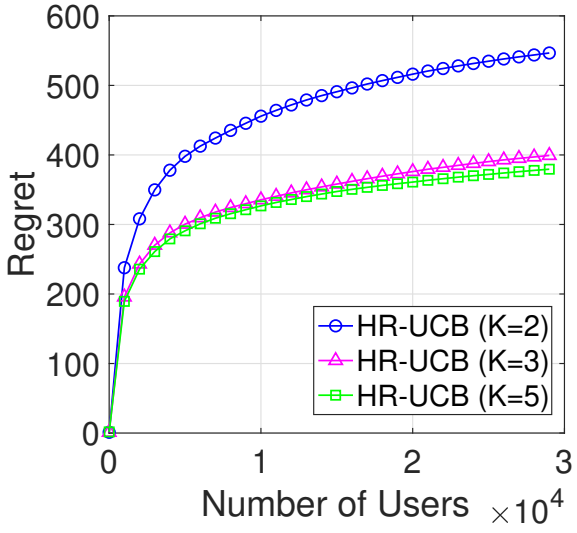
Remark 6 The assumption that β_t is known can be relaxed to the case where only the distribution of β_t is known. The analysis can be adapted to this case by (a) rewriting the reneging probability in (3.6) and $h_\beta(u,v)$ in (3.18) via integration over distribution of β_t, (b) deriving the corresponding expected lifetime under oracle policy in (3.9), and (c) reusing Theorem 1 and Lemma 5 as the GLSE does not rely on the knowledge of β_t.

Remark 7 We briefly discuss the difference between our regret bound and the regret bounds of other related settings. Note that if the satisfaction level $\beta_t = \infty$ for all t, then all the users will quit after exactly one round. This corresponds to the conventional contextual bandits setting (e.g. homoscedastic case [26] and heteroscedastic case [80]). In this degenerate case, our regret bound is $O(\sqrt{T(\log T)} \cdot \log T)$, which has an additional factor $\log T$ resulting from the heteroscedasticity.

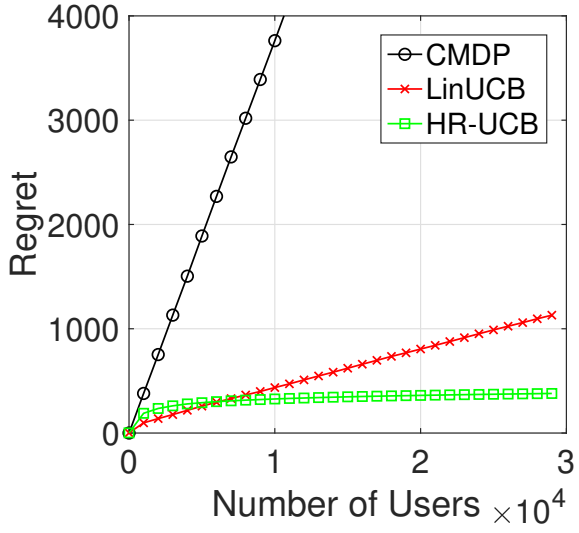
3.6 Empirical Study on the Performance of the HR-UCB Algorithm

To evaluate the empirical performance of HR-UCB, we consider 20 actions available to the decision-maker. For simplicity, the context of each user-action pair is designed to be a four-dimensional vector, which is drawn uniformly at random from a unit ball. For the mean and variance of the outcome distribution, we set $\theta_* = [0.6, 0.5, 0.5, 0.3]^\top$ and $\phi_* = [0.5, 0.2, 0.8, 0.9]^\top$, respectively. We consider the function $f(x) = x + L$ with $L = 2$ and $M_f = 1$. The acceptance level of each user is drawn uniformly at random from the interval $[-1, 1]$. We set $T = 30000$ throughout the simulations. For HR-UCB, we set $\delta = 0.1$ and $\lambda = 1$. All the results in this section are the average of 20 simulation trials. Recall that K denotes the growth rate of the regression sample set for HR-UCB. We start by evaluating the pseudo regrets of HR-UCB under different K, as shown in Figure 3.2a. Note that HR-UCB achieves a sublinear regret regardless of K. The effect of K is only reflected when the number of users is small. Specifically, a smaller K induces a slightly higher regret since it requires more users in order to accurately learn the parameters. Based on Figure 3.2a, we set $K = 5$ for the rest of the simulations.

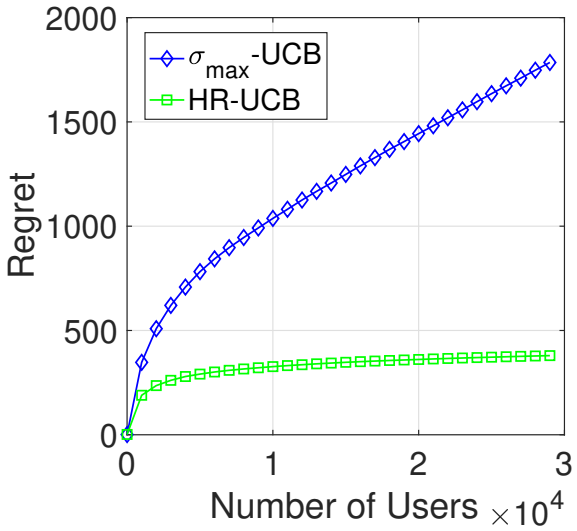
We compare the HR-UCB policy with the well-known LinUCB policy [84] and the Contextual MDP (CMDP) policy [81]. LinUCB also assumes the mean reward of arm linearly depends on the context, i.e., $\mathbb{E}[r_{t,a}|x_{t,a}] = \theta_*^\top x_{t,a}$. Different from HR-UCB, LinUCB ignores the potential dependence of the variance and the reneging risk if participants. Without planning for the reneging behavior, LinUCB always targets to maximize the cumulative rewards in an indefinite mode. In contrast, CMDP models the decision-making with sequential participants (between reneging behaviors) by episodic MDPs. At the start of each episode, the agent has access to some side-information or context that determines the dynamics of the MDP for that episode. Although CMDP



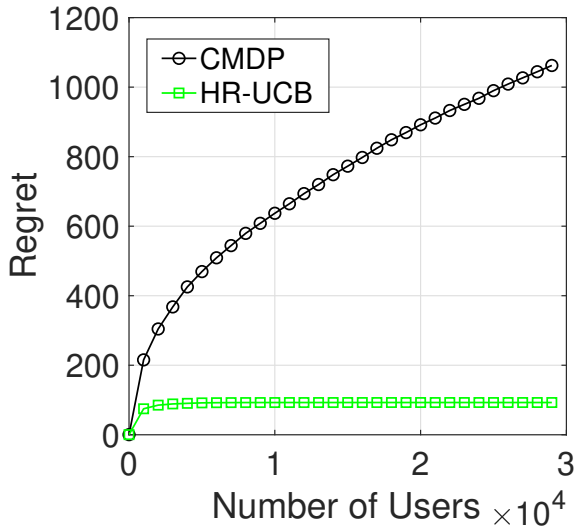
(a) Pseudo regrets: HR-UCB with different K.



(b) Pseudo regrets: LinUCB, CMDP and HR-UCB ($K = 5$).



(c) Pseudo regrets: $\sigma_{\max}$-UCB and HR-UCB ($K = 5$).



(d) Pseudo regrets: CMDP and HR-UCB ($K = 5$).

Figure 3.2: Comparison of pseudo regrets.

is able to model the reneging behaviors, it has several limitations. First, different from the HR-UCB algorithm that allows the lifetime of different participants to be random, CMDP considers that each episode has a fixed lifetime. Second, the performance of the CMDP may largely deteriorate in handling the problem handled by HR-CUB. CMDP targets to maximizing long-term rewards, which requires asymptotic optimality. In contrast, the HR-UCB algorithm aims to minimize the cumulative regret, a finer metric that requires a finite-time guarantee. Third, the scalability of CMDP is relatively poor compared to the HR-UCB algorithm. Since the space of contexts is continuous, the complexity of exploration can be very high for CMDP, especially when the dimension of features scales up.

Figure 3.2b shows the pseudo regrets under LinUCB, CMDP and HR-UCB. LinUCB achieves a linear regret because it does not take into account the heteroscedasticity of the outcome distribution in the existence of reneging. For each user, LinUCB simply chooses the action with the largest predicted mean of the outcome distribution. The regret attained by CMDP policy also appears linear. This is because CMDP handles contexts by partitioning the context space and then learning each partition-induced MDP separately. Due to the continuous context space, the CMDP policy requires numerous partitions as well as plentiful exploration for all MDPs. To make the comparison fairer, we consider a more straightforward setting with a discrete context space of size ten and only two actions (with other parameters unchanged). In this setting, Figure 3.2d shows that the regret attained by CMDP is still much larger than that by HR-UCB, and thereby shows the advantage of the proposed solution. We also consider a heuristic policy (denoted by $\sigma_{\max}$-UCB) that always assumes $\sigma_{\max}$ as the variance. We find that it tends to choose the action with the largest mean and thus incurs linear regret. We demonstrate this statement in experiments shown by Figure 3.2c, where the $\sigma_{\max}$-UCB policy attains a linear regret while HR-UCB achieves a sublinear and much smaller regret. Through simulations, we validate that HR-UCB achieves regret performance, as discussed in Section 3.4.

3.7 Possible Extensions

There are several possible directions to extend the study in this section. First, the techniques used to estimate heteroscedastic variance and establish sub-linear regret under the presence of heteroscedasticity can be extended to other variance-sensitive bandit problems, e.g., risk-averse bandits and thresholding bandits. Second, the studies can be easily adapted to another objective - maximizing total collected rewards by: (a) taking $\hat{h}_\beta(u, v) = u \cdot h_\beta(u, v)$ in regret computation, (b) reusing Theorem 1 and Lemma 5, and (c) making minor changes to constants C_3, C_4. Third, another promising extension is to use active-learning to update the sample set $\mathcal{S}$ [85]. To provide theoretical guarantees, these active-learning approaches often assume that arriving contexts are i.i.d. In contrast, since that assumption can be easily invalid (e.g., it is adversarial), we can establish the regret bound without making any such assumption. Finally, in the HR-UCB algorithm, the problem of knowledge transfer across users is given more importance than learning for a single user. This is because, compared to the population of potential users, a user's lifetime is mostly short. Therefore, another possible extension is to take into account the exploration during the lifetime of each individual user.

3.8 Summary

In this section, we propose HR-UCB – a novel learning algorithm for contextual bandits to overcome the limitation of existing bandit algorithms in applications with reneging risk and reward heteroscedasticity. Contextual bandits have been widely used to solve the sequential decision problems in many real-world applications, such as medical treatment and portfolio selection. In these applications, a "reneging" phenomenon, where participants may disengage from future interactions after an unsatisfactory outcome, is prevalent. To address the above issue, we propose a model of heteroscedastic linear bandits with reneging, which allows each participant to have a distinct "satisfaction level" with any interaction outcome falling short of that level resulting in that participant reneging. Moreover, the proposed model also allows the variance of the outcome to be context-dependent taking into account reward heteroscedasticity in real-world applications. Based

on this model, we develop the HR-UCB algorithm, and prove that it achieves $\mathcal{O}\big(\sqrt{T(\log(T))^3}\big)$ regret. We evaluate the performance of the HR-UCB algorithm by comparing its performance with baseline methods in simulation studies. The HR-UCB algorithm outperforms baseline methods under the presence of reneging risk and reward heteroscedasticity.

4. CONCLUDING REMARKS

In this dissertation, we aim to explore two fundamental challenges that are inadequately addressed in the existing literature of bandit learning. First, the efficiency of the best-performing algorithms is often unsatisfactory. Here "efficiency" is measured in terms of the performance in maximizing reward accumulation with respect to computational complexity. The gain in regret performance is often at a huge cost in computation complexity. Second, the assumptions on indefinite interaction and on reward homoscedasticity made in most existing bandit algorithms are often invalid in many real-world applications. Participants may disengage from future interactions, a phenomenon is referred to as "churn", "unsubscribing" or "reneging" in the literature. Further, rewards may be heteroscedastic, by which is meant that the variance of the reward distribution is different under different contexts. To address these challenges, we study both context-free bandits as well as contextual bandits, and propose novel learning algorithms providing theoretical guarantees on their performance. Extensive simulation experiments have been conducted to evaluate the performance of the proposed algorithms, comparing them to state-of-the-art baselines proposed algorithms are seen to outperform these baselines. We conclude by summarizing the key results as well as outlining some promising directions for future research.

- In Section 2, we study the efficiency issue in existing bandit learning algorithms and propose BMLE – a novel family of bandit algorithms. The proposed BMLE algorithms often demonstrate slightly better regret performance than other state-of-the-art bandit algorithms but with a major computational advantage. We prove that the derived BMLE indices achieve a logarithmic finite-time regret bound and hence attain order-optimality, for both exponential families and the cases beyond parametric distributions. In addition, the BMLE algorithms are formulated in a general way derived from the Biased Maximum Likelihood Estimation method that originally appeared in the adaptive control literature. They potentially enjoy great generality and thus are expected to be extendable in several promising directions, including contextual bandits, MDP,

and reinforcement learning.

- In Section 3, we study the violation of assumptions of indefinite interaction as well as reward homoscedasticity. We propose HR-UCB – a novel bandit learning algorithms to overcome the limitation of existing bandit algorithms. To address the above issue, we propose a model of heteroscedastic linear bandits with reneging, which allows each participant to have a distinct "satisfaction level" with any interaction outcome falling short of that level resulting in that participant reneging. Moreover, the proposed model also allows the variance of the outcome to be context-dependent by taking into account reward heteroscedasticity in real-world applications. We develop the HR-UCB algorithm, and prove that it achieves $\mathcal{O}\big(\sqrt{T(\log(T))^3}\big)$ regret. To the best of our knowledge, it is the first bandit algorithms to consider both the reneging risk as well as the reward heteroscedasticity. The techniques used to estimate heteroscedastic variance and establish sub-linear regret under the presence of heteroscedasticity are expected to be extendable to other variance-sensitive bandit problems, e.g., risk-averse bandits and thresholding bandits.

REFERENCES

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[2] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Conference on Learning Theory (COLT)*, pp. 217–226, 2009.

[3] R. Degenne and V. Perchet, "Anytime optimal algorithms in stochastic multi-armed bandits," in *International Conference on Machine Learning (ICML)*, pp. 1587–1595, 2016.

[4] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and Kullback-Leibler divergence," in *Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 115–122, 2010.

[5] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Conference On Learning Theory (COLT)*, pp. 359–376, 2011.

[6] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz, *et al.*, "Kullback-leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, vol. 41, no. 3, pp. 1516–1541, 2013.

[7] E. Kaufmann, O. Cappé, and A. Garivier, "On Bayesian upper confidence bounds for bandit problems," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 592–600, 2012.

[8] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2249–2257, 2011.

[9] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory (COLT)*, pp. 39–1, 2012.

[10] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: an asymptotically optimal finite-time analysis," in *International Conference on Algorithmic Learning Theory (ALT)*, pp. 199–213, 2012.

[11] N. Korda, E. Kaufmann, and R. Munos, "Thompson sampling for 1-dimensional exponential family bandits," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1448–1456, 2013.

[12] I. O. Ryzhov, W. B. Powell, and P. I. Frazier, "The knowledge gradient algorithm for a general class of online learning problems," *Operations Research*, vol. 60, no. 1, pp. 180–195, 2012.

[13] I. O. Ryzhov, P. I. Frazier, and W. B. Powell, "On the robustness of a one-period look-ahead policy in multi-armed bandit problems," *Procedia Computer Science*, vol. 1, no. 1, pp. 1635–1644, 2010.

[14] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1583–1591, 2014.

[15] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," *Operations Research*, vol. 66, no. 1, pp. 230–252, 2017.

[16] A. Towse, B. Jonsson, C. McGrath, A. Mason, R. Puig-Peiro, J. Mestre-Ferrandiz, M. Pistollato, and N. Devlin, "Understanding variations in relative effectiveness: A health production approach," *International Journal of Technology Assessment in Health Care*, vol. 31, no. 6, pp. 363–370, 2015.

[17] E. M. Buzaianu and P. Chen, "A two-stage design for comparative clinical trials: The heteroscedastic solution," *Sankhya B*, vol. 80, no. 1, pp. 151–177, 2018.

[18] C. O. Omari, P. N. Mwita, and A. W. Gichuhi, "Currency portfolio risk measurement with generalized autoregressive conditional heteroscedastic-extreme value theory-copula model," *Journal of Mathematical Finance*, vol. 8, no. 02, p. 457, 2018.

[19] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, 2003.

[20] X. Jin and T. Lehnert, "Large portfolio risk management and optimal portfolio allocation with dynamic elliptical copulas," *Dependence Modeling*, vol. 6, no. 1, pp. 19–46, 2018.

[21] X. Liu, P.-C. Hsieh, A. Bhattacharya, and P. Kumar, "Bandit learning through biased maximum likelihood estimation," *arXiv preprint arXiv:1907.01287*, 2019.

[22] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[23] J.-Y. Audibert and S. Bubeck, "Regret bounds and minimax policies under partial monitoring," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2785–2836, 2010.

[24] S. L. Scott, "A modern Bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, vol. 26, no. 6, pp. 639–658, 2010.

[25] C.-Y. Liu and L. Li, "On the prior sensitivity of Thompson sampling," in *International Conference on Algorithmic Learning Theory (ALT)*, pp. 321–336, 2016.

[26] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

[27] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2312–2320, 2011.

[28] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.

[29] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[30] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1563–1600, 2010.

[31] B. Kamiński, "Refined knowledge-gradient policy for learning probabilities," *Operations Research Letters*, vol. 43, no. 2, pp. 143–147, 2015.

[32] Y. Wang, C. Wang, and W. Powell, "The knowledge gradient for sequential decision making with stochastic binary feedbacks," in *International Conference on Machine Learning (ICML)*, pp. 1138–1147, 2016.

[33] P. R. Kumar, "A survey of some results in stochastic adaptive control," *SIAM Journal on Control and Optimization*, vol. 23, no. 3, pp. 329–380, 1985.

[34] A. A. Feldbaum, "Dual control theory. I," *Avtomatika i Telemekhanika*, vol. 21, no. 9, pp. 1240–1249, 1960.

[35] A. A. Feldbaum, "Dual control theory. II," *Avtomatika i Telemekhanika*, vol. 21, no. 11, pp. 1453–1464, 1960.

[36] V. Borkar and P. Varaiya, "Adaptive control of Markov chains, I Finite parameter set," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 953–957, 1979.

[37] V. Borkar and P. Varaiya, "Identification and adaptive control of Markov chains," *SIAM Journal on Control and Optimization*, vol. 20, no. 4, pp. 470–489, 1982.

[38] W. Lin, P. R. Kumar, and T. I. Seidman, "Will the self-tuning approach work for general cost criteria?," *Systems & control letters*, vol. 6, no. 2, pp. 77–85, 1985.

[39] A. Becker, P. R. Kumar, and C.-Z. Wei, "Adaptive control with the stochastic approximation algorithm: Geometry and convergence," *IEEE Transactions on Automatic Control*, vol. 30, no. 4, pp. 330–338, 1985.

[40] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice Hall, NJ., 1986.

[41] P. R. Kumar and A. 1, "A new family of optimal adaptive controllers for Markov chains," *IEEE Transactions on Automatic Control*, vol. 27, no. 1, pp. 137–146, 1982.

[42] P. R. Kumar and W. Lin, "Optimal adaptive controllers for unknown Markov chains," *IEEE Transactions on Automatic Control*, vol. 27, no. 4, pp. 765–774, 1982.

[43] P. R. Kumar, "Simultaneous identification and adaptive control of unknown systems over finite parameter sets," *IEEE Transactions on Automatic Control*, vol. 28, no. 1, pp. 68–76, 1983.

[44] P. R. Kumar, "Optimal adaptive control of linear-quadratic-Gaussian systems," *SIAM Journal on Control and Optimization*, vol. 21, no. 2, pp. 163–178, 1983.

[45] V. S. Borkar, "The Kumar-Becker-Lin scheme revisited," *Journal of Optimization Theory and Applications*, vol. 66, no. 2, pp. 289–309, 1990.

[46] V. S. Borkar, "Self-tuning control of diffusions without the identifiability condition," *Journal of Optimization Theory and Applications*, vol. 68, no. 1, pp. 117–138, 1991.

[47] Ł. Stettner, "On nearly self-optimizing strategies for a discrete-time uniformly ergodic adaptive model," *Applied Mathematics and Optimization*, vol. 27, no. 2, pp. 161–177, 1993.

[48] T. E. Duncan, B. Pasik-Duncan, and L. Stettner, "Almost self-optimizing strategies for the adaptive control of diffusion processes," *Journal of optimization theory and applications*, vol. 81, no. 3, pp. 479–507, 1994.

[49] M. C. Campi and P. R. Kumar, "Adaptive linear quadratic Gaussian control: the cost-biased approach revisited," *SIAM Journal on Control and Optimization*, vol. 36, no. 6, pp. 1890–1907, 1998.

[50] M. Prandini and M. C. Campi, "Adaptive LQG control of input-output systems—A cost-biased approach," *SIAM Journal on Control and Optimization*, vol. 39, no. 5, pp. 1499–1519, 2000.

[51] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press, 2019.

[52] J. Honda and A. Takemura, "Optimality of thompson sampling for gaussian bandits depends on priors," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 375–383, 2014.

[53] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is q-learning provably efficient?," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4863–4873, 2018.

[54] K. Azizzadenesheli, E. Brunskill, and A. Anandkumar, "Efficient exploration through bayesian deep q-networks," in *Information Theory and Applications Workshop (ITA)*, pp. 1–9, IEEE, 2018.

[55] N. Nikolov, J. Kirschner, F. Berkenkamp, and A. Krause, "Information-directed exploration for deep reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2019.

[56] Y. Wang, K. Dong, X. Chen, and L. Wang, "Q-learning with ucb exploration is sample efficient for infinite-horizon mdp," in *International Conference on Learning Representations (ICLR)*, 2020.

[57] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: no regret and experimental design," in *International Conference on International Conference on Machine Learning (ICML)*, pp. 1015–1022, 2010.

[58] J. Wu, M. Poloczek, A. G. Wilson, and P. Frazier, "Bayesian optimization with gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5267–5278, 2017.

[59] K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos, "Parallelised bayesian optimisation via thompson sampling," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 133–142, 2018.

[60] P.-C. Hsieh, X. Liu, A. Bhattacharya, and P. Kumar, "Stay with me: Lifetime maximization through heteroscedastic linear bandits with reneging," in *International Conference on Machine Learning (ICML)*, pp. 2800–2809, 2019.

[61] X. Liu, M. Xie, X. Wen, R. Chen, Y. Ge, N. Duffield, and N. Wang, "A semi-supervised and inductive embedding model for churn prediction of large-scale mobile games," in *IEEE International Conference on Data Mining (ICDM)*, pp. 277–286, 2018.

[62] X. Liu, M. Xie, X. Wen, R. Chen, Y. Ge, N. Duffield, and N. Wang, "Micro-and macro-level churn analysis of large-scale mobile games," *Knowledge and Information Systems*, pp. 1–32, 2019.

[63] N. McHugh, R. M. Baker, H. Mason, L. Williamson, J. van Exel, R. Deogaonkar, M. Collins, and C. Donaldson, "Extending life for people with a terminal illness: a moral right and an expensive death? exploring societal perspectives," *BMC Medical Ethics*, vol. 16, no. 1, 2015.

[64] X. Huo and F. Fu, "Risk-aware multi-armed bandit problem with application to portfolio selection," *Royal Society Open Science*, vol. 4, no. 11, p. 171377, 2017.

[65] W. Ding, T. Qiny, X.-D. Zhang, and T.-Y. Liu, "Multi-armed bandit with budget constraint and variable costs," in *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 232–238, 2013.

[66] G. Theocharous, P. S. Thomas, and M. Ghavamzadeh, "Personalized ad recommendation systems for life-time value optimization with guarantees," in *International Conference on Artificial Intelligence (IJCAI)*, pp. 1806–1812, 2015.

[67] N. Somu, G. R. MR, V. Kalpana, K. Kirthivasan, and S. S. VS, "An improved robust heteroscedastic probabilistic neural network based trust prediction approach for cloud service selection," *Neural Networks*, vol. 108, pp. 339–354, 2018.

[68] D. Niu, B. Li, and S. Zhao, "Understanding demand volatility in large vod systems," in *International Workshop on Network and Operating Systems Support for Digital Audio and Video*, pp. 39–44, 2011.

[69] R. C. H. Cheng and J. P. C. Kleijnen, "Improved design of queueing simulation experiments with highly heteroscedastic responses," *Operations Research*, vol. 47, pp. 762–777, May 1999.

[70] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3275–3283, 2012.

[71] B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier, "Qualitative multi-armed bandits: A quantile-based approach," in *International Conference on Machine Learning (ICML)*, pp. 1660–1668, 2015.

[72] A. Cassel, S. Mannor, and A. Zeevi, "A general approach to multi-armed bandits under risk criteria," in *Conference on Learning Theory (COLT)*, pp. 1295–1306, 2018.

[73] A. R. Chaudhuri and S. Kalyanakrishnan, "Quantile-regret minimisation in infinitely many-armed bandits," in *Association for Uncertainty in Artificial Intelligence (UAI)*, 2018.

[74] A. Kazerouni, M. Ghavamzadeh, Y. Abbasi, and B. Van Roy, "Conservative contextual linear bandits," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3910–3919, 2017.

[75] Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári, "Conservative bandits," in *International Conference on Machine Learning (ICML)*, pp. 1254–1262, 2016.

[76] W. Sun, D. Dey, and A. Kapoor, "Safety-aware algorithms for adversarial contextual bandit," in *International Conference on Machine Learning (ICML)*, pp. 3280–3288, 2017.

[77] J. D. Abernethy, K. Amin, and R. Zhu, "Threshold bandits, with and without censored feedback," in *Advances In Neural Information Processing Systems*, pp. 4889–4897, 2016.

[78] L. Jain and K. Jamieson, "Firing bandits: optimizing crowdfunding," in *International Conference on Machine Learning (ICML)*, pp. 2211–2219, 2018.

[79] S. Schmit and R. Johari, "Learning with abandonment," in *International Conference on Machine Learning (ICML)*, pp. 4516–4524, 2018.

[80] J. Kirschner and A. Krause, "Information directed sampling and bandits with heteroscedastic noise," in *Conference on Learning Theory (COLT)*, pp. 358–384, 2018.

[81] A. Modi, N. Jiang, S. Singh, and A. Tewari, "Markov decision processes with continuous side information," in *International Conference on Algorithmic Learning Theory (ALT)*, pp. 597–618, 2018.

[82] A. Hallak, D. Di Castro, and S. Mannor, "Contextual markov decision processes," *arXiv preprint arXiv:1502.02259*, 2015.

[83] J. M. Wooldridge, *Introductory econometrics: A modern approach*. Nelson Education, 2015.

[84] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *International Conference on World Wide Web (WWW)*, pp. 661–670, ACM, 2010.

[85] C. Riquelme, R. Johari, and B. Zhang, "Online active linear regression via thresholding," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[86] L. Erdős, H.-T. Yau, and J. Yin, "Bulk universality for generalized Wigner matrices," *Probability Theory and Related Fields*, vol. 154, no. 1-2, pp. 341–407, 2012.

[87] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[88] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

APPENDIX A

PROOFS OF SECTION 2

A.1 Proof of Lemma 1

Recall that

$$I(\nu, n, \alpha(t)) = \big(n\nu + \alpha(t)\big)\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) - n\nu\dot{F}^{-1}(\nu) - nF\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) + nF\big(\dot{F}^{-1}(\nu)\big).$$

By taking the partial derivative of $I(\nu, n, \alpha(t))$ with respect to n, we have

$$\begin{aligned}
\frac{\partial I}{\partial n} &= \nu\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) + \big(n\nu + \alpha(t)\big)\frac{\partial \dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)}{\partial n} - \nu\dot{F}^{-1}(\nu) && \text{(A.1)}\\
&\quad - F\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) - n\dot{F}\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big)\frac{\partial \dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)}{\partial n} + F\big(\dot{F}^{-1}(\nu)\big) && \text{(A.2)}\\
&= \nu \cdot \Big[\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) - \dot{F}^{-1}(\nu)\Big] - \Big[F\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) - F\big(\dot{F}^{-1}(\nu)\big)\Big]. && \text{(A.3)}
\end{aligned}$$

Since $\dot{F}(\cdot)$ is strictly increasing for the exponential families, we know $\dot{F}^{-1}(\cdot)$ is also strictly increasing and $\dot{F}^{-1}(\nu + \alpha(t)/n) > \dot{F}^{-1}(\nu)$. Moreover, by the strict convexity of $F(\cdot)$, we have

$$F\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) - F\big(\dot{F}^{-1}(\nu)\big) > \Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) - \dot{F}^{-1}(\nu)\Big) \cdot \underbrace{\dot{F}\big(\dot{F}^{-1}(\nu)\big)}_{=\nu}. \qquad \text{(A.4)}$$

Therefore, by (A.1)-(A.4), we conclude that $\frac{\partial I}{\partial n} < 0$ and hence $I(\nu, n, \alpha(t))$ is strictly decreasing with n.

A.2 Proof of Lemma 2

Recall that

$$I(\nu, n, \alpha(t)) = \big(n\nu + \alpha(t)\big)\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) - n\nu\dot{F}^{-1}(\nu) - nF\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) + nF\big(\dot{F}^{-1}(\nu)\big).$$

By taking the partial derivative of $I(\nu, n, \alpha(t))$ with respect to ν, we have

$$\frac{\partial I}{\partial \nu} = n\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) + (n\nu + \alpha(t))\frac{\partial \dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)}{\partial \nu} - \Big(n\dot{F}^{-1}(\nu) + n\nu\frac{\partial \dot{F}^{-1}(\nu)}{\partial \nu}\Big) \tag{A.5}$$

$$- n\underbrace{\dot{F}\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big)}_{\leq \nu + \alpha(t)/n}\frac{\partial \dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)}{\partial \nu} + n\underbrace{\dot{F}\Big(\dot{F}^{-1}(\nu)\Big)}_{=\nu}\frac{\partial \dot{F}^{-1}(\nu)}{\partial \nu} \tag{A.6}$$

$$\geq n \cdot \Big[\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) - \dot{F}^{-1}(\nu)\Big] > 0 \tag{A.7}$$

where the last inequality follows from the fact that $\dot{F}^{-1}(\cdot)$ is strictly increasing for the exponential families. Therefore, we can conclude that $I(\nu, n, \alpha(t))$ is strictly increasing with ν, for all $\alpha(t) > 0$ and for all $n > 0$.

A.3 Proof of Lemma 3

Recall that we define

$$\xi(k;\nu) = k\Big[\big(\nu + \frac{1}{k}\big)\dot{F}^{-1}(\nu + \frac{1}{k}) - \nu\dot{F}^{-1}(\nu)\Big] - k\Big[F\Big(\dot{F}^{-1}\big(\nu + \frac{1}{k}\big)\Big) - F(\dot{F}^{-1}(\nu))\Big], \tag{A.8}$$

$$K^*(\theta', \theta'') = \inf\{k : \dot{F}^{-1}(\theta') > \xi(k;\theta'')\}. \tag{A.9}$$

Moreover, we have $I(\mu_1, k\alpha(t), \alpha(t)) = \alpha(t)\xi(k;\mu_1)$. By Lemma 1, we know that the value of $I(\mu_1, k\alpha(t), \alpha(t))$ decreases with k, for all $k > 0$. Let $z = \frac{1}{k}$. Under any fixed $\mu_1 \in \Theta$ and $\alpha(t) > 0$, we also know that

$$\lim_{k\to\infty}\xi(k;\mu_1) = \lim_{z\downarrow 0}\frac{\big[\big(\mu_1 + z\big)\dot{F}^{-1}(\mu_1 + z) - \mu_1\dot{F}^{-1}(\mu_1)\big] - \big[F\big(\dot{F}^{-1}\big(\mu_1 + z\big)\big) - F(\dot{F}^{-1}(\mu_1))\big]}{z} \tag{A.10}$$

$$= \lim_{z\downarrow 0}\dot{F}^{-1}(\mu_1 + z) + (\mu_1 + z)\frac{\partial \dot{F}^{-1}(\mu_1 + z)}{\partial z} - \dot{F}\big(\dot{F}^{-1}(\mu_1 + z)\big)\frac{\partial \dot{F}^{-1}(\mu_1 + z)}{\partial z} \tag{A.11}$$

$$= \dot{F}^{-1}(\mu_1), \tag{A.12}$$

where (A.10) is obtained by replacing $1/k$ with z, and (A.11) follows from L'Hôpital's rule. Therefore, we have

$$\lim_{k\to\infty} I(\mu_1, k\alpha(t), \alpha(t)) = \alpha(t) \cdot \dot{F}^{-1}(\mu_1). \tag{A.13}$$

By Lemma 1 and (A.13), we know

$$I(\mu_1, k\alpha(t), \alpha(t)) \geq \alpha(t)\dot{F}^{-1}(\mu_1), \text{ for all } k > 0. \tag{A.14}$$

For any $n_2 > K^*(\mu_1, \mu_2)\alpha(t)$, we have

$$\begin{aligned} I(\mu_1, n_1, \alpha(t)) &\geq \alpha(t)\dot{F}^{-1}(\mu_1) && \text{(A.15)} \\ &\geq I(\mu_2, K^*(\mu_1, \mu_2)\alpha(t), \alpha(t)) && \text{(A.16)} \\ &> I(\mu_2, n_2, \alpha(t)), && \text{(A.17)} \end{aligned}$$

where (A.15) follows from (A.14), (A.16) holds from the definition of $K^*(\cdot, \cdot)$, and (A.17) holds due to Lemma 1. Finally, we show that $K^*(\mu_1, \mu_2)$ is finite given that $\mu_1 > \mu_2$. We consider the limit of $\xi(k; \mu_2)$ when k approaches zero and again let $z = \frac{1}{k}$:

$$\lim_{k\downarrow 0} \xi(k; \mu_2) = \lim_{z\to\infty} \frac{\left[\big(\mu_2 + z\big)\dot{F}^{-1}(\mu_2 + z) - \nu\dot{F}^{-1}(\mu_2)\right] - \left[F\big(\dot{F}^{-1}\big(\mu_2 + z\big)\big) - F(\dot{F}^{-1}(\mu_2))\right]}{z} \tag{A.18}$$

$$= \lim_{z\to\infty} \dot{F}^{-1}(\mu_2 + z) + (\mu_2 + z) \underbrace{\frac{\partial \dot{F}^{-1}(\mu_2 + z)}{\partial z}}_{\geq 0} - \underbrace{\dot{F}\big(\dot{F}^{-1}(\mu_2 + z)\big)}_{\leq \mu_2 + z} \underbrace{\frac{\partial \dot{F}^{-1}(\mu_2 + z)}{\partial z}}_{\geq 0} \tag{A.19}$$

$$\geq \lim_{z\to\infty} \dot{F}^{-1}(\mu_2 + z) \tag{A.20}$$

$$\geq \dot{F}^{-1}(\mu_1), \tag{A.21}$$

where (A.19) follows from L'Hôpital's rule and (A.21) holds due to the fact that $\dot{F}^{-1}$ is increasing.

By (A.18)-(A.21) and since $\xi(k;\mu_2)$ is continuous and strictly decreasing with k, we know there must exist a finite $k' \geq 0$ such that $\dot{F}^{-1}(\mu_1) = \xi(k';\mu_2)$. This implies that $K^*(\mu_1,\mu_2)$ is finite given that $\mu_1 > \mu_2$. □

A.4 Proof of Lemma 4

Similar to the proof of Lemma 3, we leverage the function $K^*(\cdot,\cdot)$ as defined in (A.9). By (A.9), we know that for any $k > K^*(\mu_0,\mu_2)$, we have $\xi(k;\mu_2) < \dot{F}^{-1}(\mu_0)$. Therefore, if $n_2 > K^*(\mu_0,\mu_2)\alpha(t)$,

$$I(\mu_2, n_2, \alpha(t)) < I(\mu_2, K^*(\mu_0,\mu_2), \alpha(t)) \tag{A.22}$$

$$= \alpha(t)\xi(K^*(\mu_0,\mu_2);\mu_2) \tag{A.23}$$

$$= \alpha(t)\dot{F}^{-1}(\mu_0). \tag{A.24}$$

Similarly, for any $k \leq K^*(\mu_0,\mu_1)$, we have $\xi(k;\mu_1) \geq \dot{F}^{-1}(\mu_0)$. Then, if $n_1 \leq K^*(\mu_0,\mu_1)\alpha(t)$, we know

$$I(\mu_1, n_1, \alpha(t)) \geq I(\mu_1, K^*(\mu_0,\mu_1), \alpha(t)) \tag{A.25}$$

$$= \alpha(t)\xi(K^*(\mu_0,\mu_1);\mu_1) \tag{A.26}$$

$$= \alpha(t)\dot{F}^{-1}(\mu_0). \tag{A.27}$$

Hence, by (A.22)-(A.27), we conclude that $I(\mu_1, n_1, \alpha(t)) > I(\mu_2, n_2, \alpha(t))$, for all $n_1 \leq K^*(\mu_0,\mu_1)\alpha(t)$ and $n_2 > K^*(\mu_0,\mu_2)\alpha(t)$. □

A.5 Proof of Proposition 1

Recall from (2.10) that

$$\pi_t^{\text{BMLE}} = \underset{i\in\{1,\cdots,N\}}{\operatorname{argmax}} \max_{\eta_i \in \mathcal{N}, \forall i} \Big\{ L(\mathcal{H}_t;\boldsymbol{\eta}) \exp(\eta_i \cdot \alpha(t)) \Big\}. \tag{A.28}$$

By plugging $L(\mathcal{H}_t;\boldsymbol{\eta})$ into (A.28) using the density function of the exponential families, we have

$$\pi_t^{\text{BMLE}} = \underset{i\in\{1,\cdots,N\}}{\text{argmax}} \ \underset{\eta_i\in\mathcal{N},\forall i}{\text{argmax}} \Big\{ \underbrace{\sum_{s=1}^{t} \big(\eta_{\pi_s} X_s - F(\eta_{\pi_s})\big) + \eta_i \cdot \alpha(t)}_{=:\ell_i(\mathcal{H}_t;\boldsymbol{\eta})} \Big\}. \tag{A.29}$$

Note that the inner maximization problem for $\ell_i(\mathcal{H}_t;\boldsymbol{\eta})$ over $\boldsymbol{\eta}$ is convex since $F(\cdot)$ is a convex function. Recall that $N_i(t)$ and $S_i(t)$ denote the total number of trials of arm i and the total reward collected from pulling arm i up to time t, as defined in Section 2.3. By taking the partial derivatives of $\ell_i(\mathcal{H}_t;\boldsymbol{\eta})$ with respect to each η_i, we know that $\ell_i(\mathcal{H}_t;\boldsymbol{\eta})$ is maximized when $\dot{F}(\eta_i) = \frac{S_i(t)+\alpha(t)}{N_i(t)}$ and $\dot{F}(\eta_j) = \frac{S_j(t)}{N_j(t)}$, for $j \neq i$. For each $i = 1, \cdots, N$, we then define

$$\eta_i^* := \dot{F}^{-1}\big(\frac{S_i(t)}{N_i(t)}\big), \tag{A.30}$$

$$\eta_i^{**} := \dot{F}^{-1}\big(\frac{S_i(t)+\alpha(t)}{N_i(t)}\big). \tag{A.31}$$

By plugging $\{\eta_i^*\}$ and $\{\eta_i^{**}\}$ into (A.29), we have

$$\pi_t^{\text{BMLE}} = \underset{i\in\{1,\cdots,N\}}{\text{argmax}} \Big\{ \ell_i\big(\mathcal{H}_t; \eta_i^{**}, \{\eta_j^{**}\}_{j\neq i}\big) \Big\} \tag{A.32}$$

$$= \underset{i\in\{1,\cdots,N\}}{\text{argmax}} \Big\{ \ell_i\big(\mathcal{H}_t; \eta_i^{**}, \{\eta_j^{*}\}_{j\neq i}\big) - \ell_i\big(\mathcal{H}_t; \{\eta_j^{*}\}_{j=1,\cdots,N}\big) \Big\} \tag{A.33}$$

$$= \underset{i\in\{1,\cdots,N\}}{\text{argmax}} \Big\{ \Big[\big((S_i(t)+\alpha(t))\big)\eta_i^{**} - N_i(t)F(\eta_i^{**})\Big] - \Big[S_i(t)\eta_i^{*} - N_i(t)F(\eta_i^{*})\Big] \Big\}. \tag{A.34}$$

By substituting $N_i(t)p_i(t)$ for $S_i(t)$ in (A.34), we then arrive at the index as

$$I(p_i(t), N_i(t), \alpha(t)) = \Big[\big((N_i(t)p_i(t)+\alpha(t))\big)\eta_i^{**} - N_i(t)F(\eta_i^{**})\Big] \tag{A.35}$$

$$- \Big[N_i(t)p_i(t)\eta_i^{*} - N_i(t)F(\eta_i^{*})\Big]. \tag{A.36}$$

□

A.6 Proof of Corollary 1

Recall from (A.36) that for the exponential family rewards, the BMLE index is

$$I(p_i(t), N_i(t), \alpha(t)) = \Big[\big((N_i(t)p_i(t) + \alpha(t))\big)\eta_i^{**} \tag{A.37}$$

$$- N_i(t)F(\eta_i^{**})\Big] - \Big[N_i(t)p_i(t)\eta_i^* - N_i(t)F(\eta_i^*)\Big]. \tag{A.38}$$

For the Bernoulli case, we know $F(\eta) = \log(1 + e^{\eta})$, $\dot{F}(\eta) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, $\dot{F}^{-1}(\theta) = \log(\frac{\theta}{1-\theta})$, and $F(\dot{F}^{-1}(\theta)) = \log(\frac{1}{1-\theta})$. Since $\Theta = [0, 1]$ for Bernoulli rewards, we need to analyze the following two cases when substituting the above $\dot{F}^{-1}(\theta)$ and $F(\dot{F}^{-1}(\theta))$ into (A.38):

- **Case 1:** $\alpha(t) < N_i(t)(1 - p_i(t))$ (or equivalently $\tilde{p}_i(t) < 1$)

  We have

$$I(p_i(t), N_i(t), \alpha(t)) \tag{A.39}$$

$$= \big(N_i(t)p_i(t) + \alpha(t)\big) \log\Big(\frac{N_i(t)p_i(t) + \alpha(t)}{N_i(t) - (N_i(t)p_i(t) + \alpha(t))}\Big) \tag{A.40}$$

$$- N_i(t) \log\Big(\frac{N_i(t)}{N_i(t) - (N_i(t)p_i(t) + \alpha(t))}\Big) \tag{A.41}$$

$$- N_i(t)p_i(t) \log\Big(\frac{N_i(t)p_i(t)}{N_i(t) - N_i(t)p_i(t)}\Big) + N_i(t) \log\Big(\frac{N_i(t)}{N_i(t) - N_i(t)p_i(t)}\Big) \tag{A.42}$$

$$= N_i(t)\Big\{\big(p_i(t) + \frac{\alpha(t)}{N_i(t)}\big) \log\Big(p_i(t) + \frac{\alpha(t)}{N_i(t)}\Big) \tag{A.43}$$

$$+ \big(1 - (p_i(t) + \frac{\alpha(t)}{N_i(t)})\big) \log\Big(1 - (p_i(t) + \frac{\alpha(t)}{N_i(t)})\Big) \tag{A.44}$$

$$- p_i(t) \log(p_i(t)) - (1 - p_i(t)) \log(1 - p_i(t))\Big\}, \tag{A.45}$$

  where (A.44)-(A.45) are obtained by reorganizing the terms in (A.41)-(A.42).

- **Case 2:** $\alpha(t) \geq N_i(t)(1 - p_i(t))$ (or equivalently $\tilde{p}_i(t) = 1$)

  In this case, the index would be the same as the case where $p_i(t)+\alpha(t)/N_i(t) = 1$. Therefore,

we simply have

$$I(p_i(t), N_i(t), \alpha(t)) = N_i(t)\Big\{ -p_i(t)\log(p_i(t)) - (1-p_i(t))\log(1-p_i(t))\Big\}. \quad \text{(A.46)}$$

□

A.7 Derivation of the Alternative Expression of BMLE Index in (2.17)

Note that (A.44)-(A.45) can be rewritten as follows:

$$\begin{aligned}
&I(p_i(t), N_i(t), \alpha(t)) && \text{(A.47)}\\
&= N_i(t)\Big\{ \big(p_i(t) + \frac{\alpha(t)}{N_i(t)}\big)\log\Big(p_i(t) + \frac{\alpha(t)}{N_i(t)}\Big) && \text{(A.48)}\\
&+ \big(1 - (p_i(t) + \frac{\alpha(t)}{N_i(t)})\big)\log\Big(1 - (p_i(t) + \frac{\alpha(t)}{N_i(t)})\Big) && \text{(A.49)}\\
&- p_i(t)\log(p_i(t)) - (1-p_i(t))\log(1-p_i(t))\Big\} && \text{(A.50)}\\
&= N_i(t)\Big\{ \underbrace{p_i(t)\log\Big(\frac{p_i(t) + \frac{\alpha(t)}{N_i(t)}}{p_i(t)}\Big) + (1-p_i(t))\log\Big(\frac{1 - (p_i(t) + \frac{\alpha(t)}{N_i(t)})}{1-p_i(t)}\Big)}_{=-\mathrm{KL}(p_i(t)\,||\,\tilde{p}_i(t))} \Big\} && \text{(A.51)}\\
&+ \alpha(t)\log\Big(\frac{p_i(t) + \frac{\alpha(t)}{N_i(t)}}{1 - (p_i(t) + \frac{\alpha(t)}{N_i(t)})}\Big)\\
&= \alpha(t)\log\frac{\tilde{p}_i(t)}{1-\tilde{p}_i(t)} - N_i(t)\cdot \mathrm{KL}(p_i(t)\,||\,\tilde{p}_i(t)). && \text{(A.52)}
\end{aligned}$$

□

A.8 Proof of Corollary 2

Recall from (A.36) that for the exponential family rewards, the BMLE index is

$$I(p_i(t), N_i(t), \alpha(t)) = \Big[\big((N_i(t)p_i(t) + \alpha(t))\big)\eta_i^{**} - N_i(t)F(\eta_i^{**})\Big] - \Big[N_i(t)p_i(t)\eta_i^{*} - N_i(t)F(\eta_i^{*})\Big], \quad \text{(A.53)}$$

where $\eta_i^* = \dot{F}^{-1}\big(\frac{S_i(t)}{N_i(t)}\big)$ and $\eta_i^{**} = \dot{F}^{-1}\big(\frac{S_i(t)+\alpha(t)}{N_i(t)}\big)$. For Gaussian rewards with the same variance σ^2 among arms, we have $F(\eta_i) = \sigma^2\eta_i^2/2$, $\dot{F}(\eta_i) = \sigma^2\eta_i$, $\dot{F}^{-1}(\theta_i) = \theta_i/\sigma^2$, and $F(\dot{F}^{-1}(\theta_i)) = \theta_i^2/2\sigma^2$, for each arm i. Therefore, the BMLE index becomes

$$
\begin{aligned}
&I(p_i(t), N_i(t), \alpha(t)) && \text{(A.54)}\\
&= \frac{S_i(t)+\alpha(t)}{\sigma^2 N_i(t)}(S_i(t)+\alpha(t)) - N_i(t)\frac{\sigma^2}{2}\Big(\frac{S_i(t)+\alpha(t)}{\sigma^2 N_i(t)}\Big)^2 && \text{(A.55)}\\
&- S_i(t)\frac{S_i(t)}{\sigma^2 N_i(t)} + N_i(t)\frac{\sigma^2}{2}\Big(\frac{S_i(t)}{\sigma^2 N_i(t)}\Big)^2 && \text{(A.56)}\\
&= \frac{2S_i(t)\alpha(t)+\alpha(t)^2}{2\sigma^2 N_i(t)}. && \text{(A.57)}
\end{aligned}
$$

Equivalently, for the Gaussian rewards, the selected arm at each time t is

$$
\pi_t^{\text{BMLE}} = \underset{i\in\{1,\cdots,N\}}{\operatorname{argmax}} \Big\{p_i(t) + \frac{\alpha(t)}{2N_i(t)}\Big\}. \tag{A.58}
$$

□

A.9 Proof of Corollary 3

Recall from (A.36) that for the exponential family distributions, the BMLE index is

$$
I(p_i(t), N_i(t), \alpha(t)) = \Big[\big((N_i(t)p_i(t)+\alpha(t))\big)\eta_i^{**} - N_i(t)F(\eta_i^{**})\Big] - \Big[N_i(t)p_i(t)\eta_i^* - N_i(t)F(\eta_i^*)\Big], \tag{A.59}
$$

where $\eta_i^* = \dot{F}^{-1}\big(\frac{S_i(t)}{N_i(t)}\big)$ and $\eta_i^{**} = \dot{F}^{-1}\big(\frac{S_i(t)+\alpha(t)}{N_i(t)}\big)$. For the exponential distributions, we have $F(\eta_i) = \log(\frac{-1}{\eta_i})$, $\dot{F}(\eta_i) = \frac{-1}{\eta_i}$, $\dot{F}^{-1}(\theta_i) = \frac{-1}{\theta_i}$, and $F(\dot{F}^{-1}(\theta_i)) = \log\theta_i$, for each arm i. Therefore,

the BMLE index becomes

$$I(p_i(t), N_i(t), \alpha(t)) \tag{A.60}$$

$$=(N_i(t)p_i(t)+\alpha(t))\cdot\Big(-\frac{N_i(t)}{N_i(t)p_i(t)+\alpha(t)}\Big)-N_i(t)\log\Big(\frac{N_i(t)p_i(t)+\alpha(t)}{N_i(t)}\Big) \tag{A.61}$$

$$-\Big(N_i(t)p_i(t)\big(-\frac{1}{p_i(t)}\big)\Big)+N_i(t)\log p_i(t) \tag{A.62}$$

$$=N_i(t)\log\big(\frac{N_i(t)p_i(t)}{N_i(t)p_i(t)+\alpha(t)}\big). \tag{A.63}$$

□

A.10 Proof of Proposition 2

To begin with, for each arm i, we define $p_{i,n}$ to be the empirical average reward collected in the first n pulls of arm i. For any exponential family reward distribution, the empirical mean of each arm i satisfies the following concentration inequalities [11]: For any $\delta > 0$,

$$\mathbb{P}(p_{i,n}-\theta_i \geq \delta) \leq \exp(-nD(\theta_i+\delta,\theta_i)), \tag{A.64}$$

$$\mathbb{P}(\theta_i-p_{i,n} \geq \delta) \leq \exp(-nD(\theta_i-\delta,\theta_i)). \tag{A.65}$$

Next, for each arm i, we define the following confidence intervals for each pair of $n, t \in \mathbb{N}$:

$$\delta_i^+(n,t) := \inf\Big\{\delta : \exp(-nD(\theta_i+\delta,\theta_i)) \leq \frac{1}{t^4}\Big\}, \tag{A.66}$$

$$\delta_i^-(n,t) := \inf\Big\{\delta : \exp(-nD(\theta_i-\delta,\theta_i)) \leq \frac{1}{t^4}\Big\}. \tag{A.67}$$

Accordingly, for each arm i and for each pair of $n, t \in \mathbb{N}$, we define the following events:

$$G_i^+(n,t) = \Big\{p_{i,n}-\theta_i \leq \delta_i^+(n,t)\Big\}, \tag{A.68}$$

$$G_i^-(n,t) = \Big\{\theta_i-p_{i,n} \leq \delta_i^-(n,t)\Big\}. \tag{A.69}$$

By the concentration inequality considered in Section 2.3, we have

$$\mathbb{P}(G_i^+(n,t)^{\mathsf{c}}) \le e^{-nD(\theta_i+\delta_i^+(n,t),\theta_i)} \le \frac{1}{t^4}, \tag{A.70}$$

$$\mathbb{P}(G_i^-(n,t)^{\mathsf{c}}) \le e^{-nD(\theta_i-\delta_i^-(n,t),\theta_i)} \le \frac{1}{t^4}. \tag{A.71}$$

Consider the bias term $\alpha(t) = C_\alpha \log t$ with $C_\alpha \ge 4/(D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1) \cdot K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta}))$ and $\varepsilon \in (0,1)$. Recall that we assume arm 1 is the unique optimal arm. Our target is to quantify the total number of trials of each sub-optimal arm. Define

$$Q_a(T) := \max\Big\{\frac{4}{D(\theta_a + \frac{\varepsilon}{2}\Delta_a, \theta_a)}, C_\alpha K^*(\theta_1 - \frac{\varepsilon}{2}\Delta_a, \theta_a + \frac{\varepsilon}{2}\Delta_a)\Big\}\log T + 1. \tag{A.72}$$

We start by characterizing $\mathbb{E}[N_a(T)]$ for each $a = 2, \cdots, N$:

$$\mathbb{E}[N_a(T)] \tag{A.73}$$

$$\le Q_a(T) + \mathbb{E}\Big[\sum_{t=Q_a(T)+1}^{T} \mathbb{I}\big(I(p_a(t), N_a(t), \alpha(t) \ge I(p_1(t), N_1(t), \alpha(t), N_a(t) \ge Q_a(T)\big)\Big] \tag{A.74}$$

$$= Q_a(T) + \sum_{t=Q_a(T)+1}^{T} \mathbb{P}\Big(I\big(p_a(t), N_a(t), \alpha(t)\big) \ge I\big(p_1(t), N_1(t), \alpha(t)\big), N_a(t) \ge Q_a(T)\Big) \tag{A.75}$$

$$\le Q_a(T) + \sum_{t=Q_a(T)+1}^{T} \mathbb{P}\Big(\max_{Q_a(T)\le n_a\le t} I\big(p_{a,n_a}, n_a, \alpha(t)\big) \ge \min_{1\le n_1\le t} I\big(p_{1,n_1}, n_1, \alpha(t)\big)\Big) \tag{A.76}$$

$$\le Q_a(T) + \sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t} \mathbb{P}\Big(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \ge I\big(p_{1,n_1}, n_1, \alpha(t)\big)\Big) \tag{A.77}$$

$$\le Q_a(T) + \sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t} \Big(\underbrace{\mathbb{P}\left(G_1^-(n_1,t)^{\mathsf{c}}\right)}_{\le \frac{1}{t^4}} + \underbrace{\mathbb{P}\left(G_a^+(n_a,t)^{\mathsf{c}}\right)}_{\le \frac{1}{t^4}}\Big) \tag{A.78}$$

$$+\sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t}\mathbb{P}\left(I\big(p_{a,n_a},n_a,\alpha(t)\big)\geq I\big(p_{1,n_1},n_1,\alpha(t)\big),G_1^-(n_1,t),G_a^+(n_a,t)\right) \tag{A.79}$$

$$\leq Q_a(T)+\frac{\pi^2}{3} \tag{A.80}$$

$$+\sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t}\mathbb{P}\left(I\big(p_{a,n_a},n_a,\alpha(t)\big)\geq I\big(p_{1,n_1},n_1,\alpha(t)\big),G_1^-(n_1,t),G_a^+(n_a,t)\right), \tag{A.81}$$

where the last equation follows from the fact that $\sum_{t=Q_a(T)+1}^{T}(\frac{1}{t^2})\leq \pi^2/6$. Next, to provide an upper bound for (A.81), we need to consider the following three cases separately. As suggested by (A.81), we can focus on the case where $n_a\geq Q_a(T)$.

- **Case 1:** $n_1>\frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$

  Since $n_1>\frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$, we have $p_{1,n_1}<\theta_1-\frac{\varepsilon}{2}\Delta$ on the event $G_1^-(n_1,t)$. Similarly, as $n_a\geq Q_a(T)>\frac{4}{D(\theta_a+\frac{\varepsilon}{2}\Delta_a,\theta_a)}\log t$, we have $p_{a,n_a}\leq\theta_a+\frac{\varepsilon}{2}\Delta_a$ on the event $G_a^+(n_a,t)$. Therefore, we know

  $$p_{1,n_1}-p_{a,n_a}>(1-\varepsilon)\Delta. \tag{A.82}$$

  Then, we have

  $$I(p_{1,n_1},n_1,\alpha(t))>I(\theta_1-\frac{\varepsilon}{2}\Delta,n_1,\alpha(t)) \tag{A.83}$$

  $$\geq I(\theta_a-\frac{\varepsilon}{2}\Delta,K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_a+\frac{\varepsilon}{2}\Delta)\alpha(t),\alpha(t)) \tag{A.84}$$

  $$\geq I(\theta_a-\frac{\varepsilon}{2}\Delta_a,K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_a+\frac{\varepsilon}{2}\Delta)\alpha(t),\alpha(t)) \tag{A.85}$$

  $$\geq I(p_{a,n_a},K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_a+\frac{\varepsilon}{2}\Delta)\alpha(t),\alpha(t)) \tag{A.86}$$

  $$\geq I(p_{a,n_a},Q_a(T),\alpha(t)) \tag{A.87}$$

  $$\geq I(p_{a,n_a},n_a,\alpha(t)), \tag{A.88}$$

  where (A.83) and (A.85)-(A.86) hold by Lemma 2, (A.84) holds by Lemma 3, and (A.87)-

(A.88) follow from Lemma 1. Hence, in Case 1, we always have $I\big(p_{1,n_1}, n_1, \alpha(t)\big) > I\big(p_{a,n_a}, n_a, \alpha(t)\big)$.

- **Case 2:** $n_1 \leq \frac{4}{D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1)} \log t$ and $n_1 \leq K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta})\alpha(t)$

  Similar to Case 1, since $n_a \geq Q_a(T) > \frac{4}{D(\theta_a, \theta_a + \frac{\varepsilon}{2}\Delta_a)} \log t$, we have $p_{a,n_a} \leq \theta_a + \frac{\varepsilon}{2}\Delta_a$ on the event $G_a^+(n_a, t)$. Moreover, as $n_1 \leq K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta})\alpha(t)$ and $n_a \geq Q_a(T) > K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_a + \frac{\varepsilon}{2}\Delta)\alpha(t)$, by Lemma 4 we know

$$I(\underline{\theta}, n_1, \alpha(t)) > I(\theta_a + \frac{\varepsilon}{2}\Delta, n_a, \alpha(t)). \tag{A.89}$$

  Therefore, we obtain that

$$\begin{aligned} I\big(p_{1,n_1}, n_1, \alpha(t)\big) &> I\big(\underline{\theta}, n_1, \alpha(t)\big) && \text{(A.90)} \\ &> I(\theta_a + \frac{\varepsilon}{2}\Delta, n_a, \alpha(t)) && \text{(A.91)} \\ &> I\big(p_{a,n_a}, n_a, \alpha(t)\big), && \text{(A.92)} \end{aligned}$$

  where (A.90) and (A.92) follow from Lemma 2, and (A.91) is a direct result of (A.89). Hence, in Case 2, we still have $I\big(p_{1,n_1}, n_1, \alpha(t)\big) > I\big(p_{a,n_a}, n_a, \alpha(t)\big)$.

- **Case 3:** $n_1 \leq \frac{4}{D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1)} \log t$ and $n_1 > K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta})\alpha(t)$

  Recall that $\alpha(t) = C_\alpha \log t$ with $C_\alpha \geq 4/(D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1) \cdot K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta}))$. Therefore, the two events $\{n_1 \leq \frac{4}{D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1)} \log t\}$ and $\{n_1 > K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta})\alpha(t)\}$ cannot happen at the same time.

To sum up, in all the above three cases, we have

$$\mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1^-(n_1, t), G_a^+(n_a, t)\right) = 0. \tag{A.93}$$

By (A.81) and (A.93), we conclude that $E[N_a(T)] \leq Q_a(T) + \frac{\pi^2}{3}$, for every $a \neq 1$.

Finally, the total regret can be upper bounded as

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \cdot E[N_a(T)] \tag{A.94}$$

$$= \sum_{a=2}^{N} \Delta_a \Bigg[\max\Big\{ \frac{4}{D(\theta_a + \frac{\varepsilon}{2}\Delta_a, \theta_a)}, C_\alpha K^*(\theta_1 - \frac{\varepsilon}{2}\Delta_a, \theta_a + \frac{\varepsilon}{2}\Delta_a) \Big\} \log T + 1 + \frac{\pi^2}{3} \Bigg]. \tag{A.95}$$

□

A.11 Proof of Proposition 3

We extend the proof of Proposition 2 to the case of Gaussian rewards. To begin with, we define the confidence intervals and the "good" events. Recall that for each arm i, we define $p_{i,n}$ to be the empirical average reward collected in the first n pulls of arm i. For each arm i, for each pair of $n, t \in \mathbb{N}$, we define

$$\delta_i(n,t) := \inf\Big\{ \delta : \max\big\{ \exp(-nD(\theta_i + \delta, \theta_i)), \exp(-nD(\theta_i - \delta, \theta_i)) \big\} \leq \frac{1}{t^4} \Big\}. \tag{A.96}$$

Accordingly, for each arm i and for each pair of $n, t \in \mathbb{N}$, we define the following events:

$$G_i(n,t) = \Big\{ |p_{i,n} - \theta_i| \leq \delta_i(n,t) \Big\}, \tag{A.97}$$

For the Gaussian rewards, we can leverage Hoeffding's inequality for sub-Gaussian distributions as follows:

Lemma 11. *Under σ-sub-Gaussian rewards for all arms, for any $n \in \mathbb{N}$, we have*

$$\mathbb{P}(|p_{i,n} - \theta_i| \geq \delta) \leq 2\exp(-\frac{n}{2\sigma^2}\delta^2). \tag{A.98}$$

The proof of Lemma 11 is a direct result of Proposition 2.5 in [51]. □

Based on Lemma 11, we shall focus on the case $D(\theta', \theta'') = \frac{1}{2\sigma^2}(|\theta' - \theta''|)^2$ and $\delta_i(n,t) =$

$\sqrt{(8\sigma^2 \log t)/n}$. For ease of notation, we use γ_* to denote the constant $8\sigma^2$.

Before providing the regret analysis, we first introduce the following useful lemma.

Lemma 12. *Suppose $\gamma > 0$ and $\mu_1, \mu_2 \in \mathbb{R}$ with $\mu_1 > \mu_2$. Given $\alpha(t) = c \log t$ with $c \geq \frac{32\gamma}{\mu_1 - \mu_2}$, for any $n_2 \geq \frac{2}{\mu_1 - \mu_2}\alpha(t)$ and any $n_1 > 0$, we have $I(\mu_1 - \sqrt{(\gamma \log t)/n_1}, n_1, \alpha(t)) > I(\mu_2 + \sqrt{(\gamma \log t)/n_2}, n_2, \alpha(t))$.*

The proof of Lemma 12 is summarized as below. We start by considering $n_2 \geq M\alpha(t)$, for some $M > 0$. Then, note that

$$I\Big(\mu_1 - \sqrt{\frac{\gamma \log t}{n_1}}, n_1, \alpha(t)\Big) = \mu_1 - \sqrt{\frac{\gamma \log t}{n_1}} + \frac{\alpha(t)}{2n_1}, \tag{A.99}$$

$$I\Big(\mu_2 + \sqrt{\frac{\gamma \log t}{n_2}}, n_2, \alpha(t)\Big) = \mu_2 - \sqrt{\frac{\gamma \log t}{n_2}} + \frac{\alpha(t)}{2n_2}, \tag{A.100}$$

For ease of notation, we use x_1 and x_2 to denote $\sqrt{(\gamma \log t)/n_1}$ and $\sqrt{(\gamma \log t)/n_2}$, respectively. Then, we know

$$I\Big(\mu_1 - \sqrt{\frac{\gamma \log t}{n_1}}, n_1, \alpha(t)\Big) - I\Big(\mu_2 + \sqrt{\frac{\gamma \log t}{n_1}}, n_2, \alpha(t)\Big) \tag{A.101}$$

$$\geq (\mu_1 - \mu_2) - (x_1 + x_2) + \frac{c}{2\gamma}(x_1^2 - x_2^2) \tag{A.102}$$

$$\geq (\mu_1 - \mu_2) - x_1 - \sqrt{\frac{\gamma}{cM}} + \frac{c}{2\gamma}x_1^2 - \frac{1}{2M}, \tag{A.103}$$

where (A.103) follows from that $n_2 \geq M\alpha(t)$. Define $w(x_1) := (\mu_1 - \mu_2) - x_1 - \sqrt{\frac{\gamma}{cM}} + \frac{c}{2\gamma}x_1^2 - \frac{1}{2M}$. The quadratic polynomial $w(x_1)$ remains positive for all $x_1 \in \mathbb{R}$ if the discriminant of $w(x_1)$, denoted by Disc$(w(x_1))$, is negative. Indeed, we have

$$\text{Disc}(w(x_1)) = 1 - 4 \cdot \frac{c}{2\gamma} \cdot (-\sqrt{\frac{\gamma}{cM}} - \frac{1}{2M} + (\mu_1 - \mu_2)) \leq -39, \tag{A.104}$$

where the last inequality follows from that $c \geq \frac{32\gamma}{\mu_1 - \mu_2}$ and $M = \frac{2}{\mu_1 - \mu_2}$. □

Now, we are ready to prove Proposition 3: Consider the bias term $\alpha(t) = C_\alpha \log t$ with $C_\alpha \geq \frac{32\gamma^*}{\Delta}$, where $\gamma^* = 8\sigma^2$. Recall that we assume arm 1 is the unique optimal arm. Our target is to

quantify the total number of trials of each sub-optimal arm. Next, we characterize the expected total number of trials of each sub-optimal arm, i.e. $\mathbb{E}[N_a(T)]$. We define $Q_a^*(T) = \frac{2}{\Delta_a} C_\alpha \log T$. By using as similar argument to (A.73)-(A.81), we have

$$\mathbb{E}[N_a(T)] \leq Q_a^*(T) \tag{A.105}$$

$$+ \sum_{t=Q_a(T)+1}^{T} \mathbb{P}\left(I\big(p_a(t), N_a(t), \alpha(t)\big) \geq I\big(p_1(t), N_1(t), \alpha(t)\big), N_a(t) \geq Q_a^*(T) \right) \tag{A.106}$$

$$\leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big) \right) \tag{A.107}$$

$$\leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \Big(\underbrace{\mathbb{P}\left(G_1(n_1,t)^{\mathsf{c}}\right)}_{\leq \frac{2}{t^4}} + \underbrace{\mathbb{P}\left(G_a(n_a,t)^{\mathsf{c}}\right)}_{\leq \frac{2}{t^4}} \Big) \tag{A.108}$$

$$+ \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1(n_1,t), G_a(n_a,t) \right) \tag{A.109}$$

$$\leq Q_a^*(T) + \frac{2\pi^2}{3} \tag{A.110}$$

$$+ \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1(n_1,t), G_a(n_a,t) \right). \tag{A.111}$$

Conditioned on the events $G_i(n_1,t)$ and $G_a(n_a,t)$, we obtain that

$$I\big(p_{1,n_1}, n_1, \alpha(t)\big) \geq I(\theta_1 - \sqrt{(\gamma_* \log t)/n_1}, n_1, \alpha(t)) \tag{A.112}$$

$$> I(\theta_a + \sqrt{(\gamma_* \log t)/n_a}, n_a, \alpha(t)) \tag{A.113}$$

$$\geq I\big(p_{a,n_a}, n_a, \alpha(t)\big), \tag{A.114}$$

where (A.112) and (A.114) follow from Lemma 2, and (A.113) follows from Lemma 12. Hence,

for $n_1 > 0$ and $n_a \geq Q_a^*(T)$,

$$\mathbb{P}\Big(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1(n_1, t), G_a(n_a, t)\Big) = 0. \tag{A.115}$$

By (A.111) and (A.115), we know $E[N_a(T)] \leq Q_a^*(T) + \frac{2\pi^2}{3}$, for every $a \neq 1$. Hence, the total regret can be upper bounded as

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[\frac{2}{\Delta_a} C_\alpha \log T + \frac{2\pi^2}{3}\big]. \tag{A.116}$$

□

A.12 Proof of Proposition 5

For sub-exponential reward distributions, we consider the sub-exponential tail bound as follows:

Lemma 13. *Under (ρ, κ)-sub-exponential rewards for all arms, for any $n \in \mathbb{N}$, we have*

$$\mathbb{P}(p_{i,n} - \theta_i \geq \delta) \leq \exp\Big(-\frac{n^2\delta^2}{2(n\kappa\delta + \rho^2)}\Big). \tag{A.117}$$

Similar to the proof of Proposition 2, we consider the bias term $\alpha(t) = C_\alpha \log t$, but with $C_\alpha \geq 16(\kappa\varepsilon\Delta + 2\rho^2)/((\varepsilon\Delta)^2 K^*(\theta_1 - \frac{\varepsilon\Delta}{2}, 0))$. Note that here we simply replace $D(\theta_1 - \frac{\varepsilon\Delta}{2}, \theta_1)$ with $\frac{(\varepsilon\Delta)^2}{4(\kappa\varepsilon\Delta + 2\rho^2)}$ by comparing (A.117) with (A.64). Similarly, we define

$$\tilde{Q}_a(T) := \max\Big\{\frac{16(\kappa\varepsilon\Delta + 2\rho^2)}{(\varepsilon\Delta_a)^2}, C_\alpha K^*(\theta_1 - \frac{\varepsilon}{2}\Delta_a, \theta_a + \frac{\varepsilon}{2}\Delta_a)\Big\} \log T + 1. \tag{A.118}$$

Note that the proof of Proposition 2 relies only on Lemmas 1-4, and these lemmas are tied to the distributions for deriving the BMLE index, not to the underlying true reward distributions. Therefore, it is easy to verify that the same proof procedure still holds here by replacing $Q_a(T)$ with $\tilde{Q}_a(T)$. □

APPENDIX B

PROOFS OF SECTION 3

B.1 Proof of Lemma 6

Proof. Recall that $\boldsymbol{V}_n = \big(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}_d\big)$. Note that

$$\widehat{\phi}_n = (\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda I_d)^{-1} \boldsymbol{X}_n^\top f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) \tag{B.1}$$

$$= \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^\top f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) \tag{B.2}$$

$$= \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^\top \big(f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) - \boldsymbol{X}_n \phi_* + \boldsymbol{X}_n \phi_*\big) \tag{B.3}$$

$$+ \lambda \boldsymbol{V}_n^{-1} \phi_* - \lambda \boldsymbol{V}_n^{-1} \phi_* \tag{B.4}$$

$$= \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^\top \big(f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) - \boldsymbol{X}_n \phi_*\big) - \lambda \boldsymbol{V}_n^{-1} \phi_* + \phi_*. \tag{B.5}$$

Therefore, for any $x \in \mathbb{R}^d$, we know

$$|x^\top \widehat{\phi}_n - x^\top \widehat{\phi}_*| \tag{B.6}$$

$$= |x^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^\top \big(f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) - \boldsymbol{X}_n \phi_*\big) - \lambda x^\top \boldsymbol{V}_n^{-1} \phi_*| \tag{B.7}$$

$$\leq \|x\|_{\boldsymbol{V}_n^{-1}} \Big(\lambda \|\phi_*\|_{\boldsymbol{V}_n^{-1}} \tag{B.8}$$

$$+ \big\|\boldsymbol{X}_n^\top \big(f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) - \boldsymbol{X}_n \phi_*)\big)\big\|_{\boldsymbol{V}_n^{-1}} \Big). \tag{B.9}$$

Moreover, by rewriting $\widehat{\varepsilon} = \widehat{\varepsilon} - \varepsilon + \varepsilon$, we have

$$f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) \tag{B.10}$$

$$= f^{-1}\big((\widehat{\varepsilon} - \varepsilon + \varepsilon) \circ (\widehat{\varepsilon} - \varepsilon + \varepsilon)\big) \tag{B.11}$$

$$= f^{-1}(\varepsilon \circ \varepsilon) + M_f^{-1}\Big(2\big(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\big) \tag{B.12}$$

$$+ \big(\boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\big)\Big), \tag{B.13}$$

where (B.12)-(B.13) follow from the fact that both $f(\cdot)$ and $f^{-1}(\cdot)$ are linear with a slope M_f and M_f^{-1}, respectively, as described in Section 3.3. Therefore, by (B.6)-(B.13) and the Cauchy-Schwarz inequality, we have

$$|x^\top \widehat{\phi}_n - x^\top \widehat{\phi}_*| \leq \|x\|_{\boldsymbol{V}_n{}^{-1}} \Big\{ \lambda \|\phi_*\|_{\boldsymbol{V}_n^{-1}} \tag{B.14}$$

$$+ \big\| \boldsymbol{X}_n^\top \big(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n \phi_* \big) \big\|_{\boldsymbol{V}_n{}^{-1}} \tag{B.15}$$

$$+ 2M_f^{-1} \Big\| \boldsymbol{X}_n^\top \big(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \big) \Big\|_{\boldsymbol{V}_n{}^{-1}} \tag{B.16}$$

$$+ M_f^{-1} \Big\| \boldsymbol{X}_n^\top \big(\boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \big) \Big\|_{\boldsymbol{V}_n{}^{-1}} \Big\}. \tag{B.17}$$

□

B.2 Proof of Lemma 7

We first introduce the following useful lemmas.

Lemma 14 (Lemma 8.2 in [86])**.** *Let $\{a_i\}_{i=1}^N$ be N independent random complex variables with zero mean and variance σ^2 and having uniform sub-exponential decay, i.e., there exists $\kappa_1, \kappa_2 > 0$ such that*

$$\mathbb{P}\{|a_i| \geq x^{\kappa_1}\} \leq \kappa_2 e^{-x}. \tag{B.18}$$

We use a^{H} to denote the conjugate transpose of a. Let $a = (a_1, \cdots, a_N)^\top$, let $\overline{a_i}$ denote the complex conjugate of a_i, for all i, and let $\boldsymbol{B} = (B_{ij})$ be a complex $N \times N$ matrix. Then, we have

$$\mathbb{P}\Big\{ |a^{\mathrm{H}} \boldsymbol{B} a - \sigma^2 \mathrm{tr}(\boldsymbol{B})| \geq s\sigma^2 \Big(\sum_{i=1}^N |B_{ii}|^2 \Big)^{-1/2} \Big\} \tag{B.19}$$

$$\leq C_1 \exp\Big(- C_2 \cdot s^{1/(1+\kappa_1)} \Big), \tag{B.20}$$

where C_1 and C_2 are positive constants that depend only on κ_1, κ_2. Moreover, for the standard χ_1^2-distribution, $\kappa_1 = 1$ and $\kappa_2 = 2$.

For any $p \times q$ matrix $\boldsymbol{A}$, we define the induced matrix norm as $\|\boldsymbol{A}\|_2 := \max_{v \in \mathbb{R}^q, \|v\|_2 = 1} \|\boldsymbol{A}v\|_2$.

Lemma 15.

$$\left\| \boldsymbol{V}_n^{-1/2} \boldsymbol{X}^\top \right\|_2 \leq 1, \forall n \in \mathbb{N}. \tag{B.21}$$

Proof. By the definition of induced matrix norm,

$$\begin{aligned}
\left\| \boldsymbol{V}_n^{-1/2} \boldsymbol{X}^\top \right\|_2 &= \max_{\|v\|_2 = 1} \sqrt{v^\top \boldsymbol{X} \boldsymbol{V}_n^{-1} \boldsymbol{X}^\top v} && \text{(B.22)} \\
&= \lambda_{\max}\Big(\boldsymbol{X} \boldsymbol{V}_n^{-1} \boldsymbol{X}^T \Big) && \text{(B.23)} \\
&= \lambda_{\max}\Big(\boldsymbol{X} \big(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_d \big)^{-1} \boldsymbol{X}^T \Big) && \text{(B.24)} \\
&\leq \frac{\lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X})}{\lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) + \lambda} \leq 1, && \text{(B.25)}
\end{aligned}$$

where (B.25) follows from the singular value decomposition and $\lambda_{\max}(\boldsymbol{X}^\top \boldsymbol{X}) \geq 0$. □

To simplify notation, we use $\boldsymbol{X}$ and $\boldsymbol{V}$ as a shorthand for $\boldsymbol{X}_n$ and $\boldsymbol{V}_n$, respectively. For convenience, we rewrite $\boldsymbol{V}^{-1/2} X^\top = [v_1 \cdots v_n]$ as the matrix of n column vectors $\{v_i\}_{i=1}^n$ (each $v_i \in \mathbb{R}^d$) and show the following property.

Lemma 16. *Let $v_i \in \mathbb{R}^d$ be the i-th column of the matrix $\boldsymbol{V}^{-1/2} X^\top$, for all $1 \leq i \leq n$. Then, we have*

$$\sum_{i=1}^n \|v_i\|_2^2 \leq d. \tag{B.26}$$

Proof of Lemma 16. Recall that $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a square matrix. We know

$$\begin{aligned}
\sum_{i=1}^n \|v_i\|_2^2 &= \mathrm{tr}\Big(\big(\boldsymbol{X} \boldsymbol{V}^{-1/2} \big) \big(\boldsymbol{V}^{-1/2} \boldsymbol{X}^\top \big) \Big) && \text{(B.27)} \\
&= \mathrm{tr}\Big(\big(\boldsymbol{V}^{-1/2} \boldsymbol{X} \big) \big(\boldsymbol{X}^\top \boldsymbol{V}^{-1/2} \big) \Big) && \text{(B.28)} \\
&\leq d \cdot \lambda_{\max}\Big(\big(\boldsymbol{V}^{-1/2} \boldsymbol{X} \big) \big(\boldsymbol{X}^\top \boldsymbol{V}^{-1/2} \big) \Big), && \text{(B.29)}
\end{aligned}$$

where (B.28) follows from the trace of a product being commutative, and (B.29) follows since the trace is the sum of all eigenvalues. Moreover, we have

$$\lambda_{\max}\big((\boldsymbol{X}\boldsymbol{V}^{1/2})(\boldsymbol{X}^\top\boldsymbol{V}^{-1/2})\big) \tag{B.30}$$

$$= \big\|(\boldsymbol{X}\boldsymbol{V}^{1/2})(\boldsymbol{X}^\top\boldsymbol{V}^{-1/2})\big\|_2 \tag{B.31}$$

$$\leq \big\|(\boldsymbol{X}\boldsymbol{V}^{1/2})\big\|_2 \big\|(\boldsymbol{X}^\top\boldsymbol{V}^{-1/2})\big\|_2 \leq 1, \tag{B.32}$$

where (B.32) follows from the fact that the ℓ_2-norm is sub-multiplicative. Therefore, by (B.27)-(B.32), we conclude that $\sum_{i=1}^n \|v_i\|_2^2 \leq d$. □

We are now ready to prove Lemma 7.

Proof of Lemma 7. To simplify notation, we use $\boldsymbol{X}$ and $\boldsymbol{V}$ as a shorthand for $\boldsymbol{X}_n$ and $\boldsymbol{V}_n$, respectively. To begin with, we know $f^{-1}(\varepsilon\circ\varepsilon) - \boldsymbol{X}\phi_* = \frac{1}{M_f}((\varepsilon\circ\varepsilon) - f(\boldsymbol{X}\phi_*))$. Therefore, we have

$$\big\|\boldsymbol{X}(f^{-1}(\varepsilon\circ\varepsilon) - \boldsymbol{X}\phi_*)\big\|_{\boldsymbol{V}^{-1}} \tag{B.33}$$

$$= \frac{1}{M_f}\sqrt{\big(\varepsilon\circ\varepsilon - f(\boldsymbol{X}\phi_*)\big)^\top \boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top\big(\varepsilon\circ\varepsilon - f(\boldsymbol{X}\phi_*)\big)}, \tag{B.34}$$

where each element in the vector $(\varepsilon\circ\varepsilon - f(\boldsymbol{X}\phi_*))$ is a centered χ_1^2-distribution with a scaling of $f(\phi_*^\top x_i)$. Defining $\boldsymbol{W} = \mathrm{diag}\big(f(x_1^\top\phi_*), ..., f(x_n^\top\phi_*)\big)$, we have

$$\big\|\boldsymbol{X}(f^{-1}(\varepsilon\circ\varepsilon) - \boldsymbol{X}\phi_*)\big\|_{\boldsymbol{V}^{-1}} \tag{B.35}$$

$$= \frac{1}{M_f}\Big[\underbrace{\Big(\varepsilon\circ\varepsilon - f(\boldsymbol{X}\phi_*)\Big)^\top \boldsymbol{W}^{-1}}_{\text{mean=0, variance= 2}}\Big(\boldsymbol{W}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top\boldsymbol{W}\Big) \tag{B.36}$$

$$\underbrace{\boldsymbol{W}^{-1}\Big(\varepsilon\circ\varepsilon - f(\boldsymbol{X}\phi_*)\Big)}_{\text{mean=0, variance=2}}\Big]^{1/2}. \tag{B.37}$$

We use $\eta = \boldsymbol{W}^{-1}\big(\varepsilon\circ\varepsilon - f(\boldsymbol{X}\phi_*)\big)$ as a shorthand and define $\boldsymbol{U} = \big(U_{ij}\big) = \boldsymbol{W}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top\boldsymbol{W}$. By Lemma 14 and the fact that $\varepsilon(x_1), \cdots, \varepsilon(x_n)$ are mutually independent given the contexts $\{x_i\}_{i=1}^n$,

we have

$$\mathbb{P}\Big\{|\eta^\top \boldsymbol{U}\eta - 2\cdot \text{tr}(\boldsymbol{U})| \geq 2s\Big(\sum_{i=1}^{n}|\boldsymbol{U}_{ii}|^2\Big)^{1/2}\Big\} \tag{B.38}$$

$$\leq C_1 \text{exp}(-C_2\sqrt{s}). \tag{B.39}$$

Recall that $\boldsymbol{V}^{-1/2}X^\top = [v_1 \cdots v_n]$. The trace of $\boldsymbol{U}$ can be upper bounded as

$$\text{tr}(\boldsymbol{U}) = \text{tr}(\boldsymbol{W}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top\boldsymbol{W}) \tag{B.40}$$

$$= \text{tr}\Big(\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{W}\boldsymbol{X}\boldsymbol{V}^{-1/2}\Big) \tag{B.41}$$

$$= \sum_{i=1}^{n} f(x_i^\top\phi_*)^2 \cdot \|v_i\|_2^2 \tag{B.42}$$

$$\leq (\sigma_{\max}^2)^2 \sum_{i=1}^{n}\|v_i\|_2^2 \leq (\sigma_{\max}^2)^2 d, \tag{B.43}$$

where the last inequality in (B.43) follows directly from Lemma 16. Also by the commutative property of the trace operation, we have

$$\sum_{i=1}^{n}|\boldsymbol{U}_{ii}|^2 \overset{(a)}{\leq} \Big(\sum_{i=1}^{n}\boldsymbol{U}_{ii}\Big)^2 \overset{(b)}{\leq} \big((\sigma_{\max}^2)^2 d\big)^2, \tag{B.44}$$

where (a) follows from $\boldsymbol{U}$ being positive semi-definite (all diagonal elements are nonnegative), and (b) follows from (B.43). Therefore, by (B.38)-(B.44), we have

$$\mathbb{P}\Big\{\eta^\top\boldsymbol{U}\eta \geq 2s\cdot(\sigma_{\max}^2)^2 d + 2(\sigma_{\max}^2)^2 d\Big\} \tag{B.45}$$

$$\leq C_1 \cdot \text{exp}(-C_2\sqrt{s}). \tag{B.46}$$

By choosing $s = \Big(\frac{1}{C_2}\ln\frac{C_1}{\delta}\Big)^2$, we have

$$\mathbb{P}\Big\{\eta^\top\boldsymbol{U}\eta \geq 2(\sigma_{\max}^2)^2 d\Big(\Big(\frac{1}{C_2}\ln\frac{C_1}{\delta}\Big)^2 + 1\Big)\Big\} \leq \delta. \tag{B.47}$$

Therefore, we conclude that with probability at least $1-\delta$, the following inequality holds

$$\left\|\boldsymbol{X}(f^{-1}(\varepsilon\circ\varepsilon)-\boldsymbol{X}\phi_*)\right\|_{\boldsymbol{V}^{-1}} \tag{B.48}$$

$$\leq \frac{1}{M_f}\sqrt{2(\sigma_{\max}^2)^2\cdot d\Big(\Big(\frac{1}{C_2}\ln\frac{C_1}{\delta}\Big)^2+1\Big)}. \tag{B.49}$$

□

B.3 Proof of Lemma 8

We first introduce a useful lemma.

Lemma 17 (Theorem 4.1 in [87])**.** *Consider a finite sequence $\{\boldsymbol{A}_k\}$ of fixed self-adjoint matrices of dimension $d\times d$, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Let $\sigma^2=\|\sum_k \boldsymbol{A}_k^2\|_2$. Then, for all $s\geq 0$,*

$$\mathbb{P}\Big\{\lambda_{\max}\Big(\sum_k\gamma_k\boldsymbol{A}_k\Big)\geq s\Big\}\leq d\cdot\exp(-\frac{s^2}{2\sigma^2}), \tag{B.50}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a square matrix.

Now we are ready to prove Lemma 8.

Proof of Lemma 8. To simplify notation, we use $\boldsymbol{X}$ and $\boldsymbol{V}$ as a shorthand for $\boldsymbol{X}_n$ and $\boldsymbol{V}_n$, respectively. Recall that $\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top=[v_1,v_2,...,v_n]$ and define $\boldsymbol{A}_i=v_iv_i^\top$, for all $i=1,...,n$. Note that $\boldsymbol{A}_i$ is symmetric, for all i. Define an $n\times n$ diagonal matrix $\boldsymbol{D}=\mathrm{diag}(\varepsilon_1,\varepsilon_2,...,\varepsilon_n)$. Then we have:

$$\Big\|\boldsymbol{X}^\top\Big(\varepsilon\circ\big(\boldsymbol{X}(\theta_*-\widehat{\theta})\big)\Big)\Big\|_{\boldsymbol{V}^{-1}} \tag{B.51}$$

$$=\Big\|\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top\Big(\varepsilon\circ\big(\boldsymbol{X}(\theta_*-\widehat{\theta})\big)\Big)\Big\|_2 \tag{B.52}$$

$$=\Big\|\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{X}(\theta_*-\widehat{\theta})\Big\|_2 \tag{B.53}$$

$$\tag{B.54}$$

$$= \left\| \boldsymbol{V}^{-1/2} \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X} \boldsymbol{V}^{-1/2} \boldsymbol{V}^{1/2} (\theta_* - \widehat{\theta}) \right\|_2 \tag{B.55}$$

$$\leq \left\| \boldsymbol{V}^{-1/2} \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X} \boldsymbol{V}^{-1/2} \right\|_2 \cdot \left\| \boldsymbol{V}^{1/2} (\theta_* - \widehat{\theta}) \right\|_2 \tag{B.56}$$

$$= \left\| \boldsymbol{V}^{-1/2} \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X} \boldsymbol{V}^{-1/2} \right\|_2 \cdot \left\| \theta_* - \widehat{\theta} \right\|_{\boldsymbol{V}}. \tag{B.57}$$

Next, the first term in (B.57) can be expanded into

$$\left\| \boldsymbol{V}^{-1/2} \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X} \boldsymbol{V}^{-1/2} \right\|_2 \tag{B.58}$$

$$= \left\| \sum_{i=1}^n \varepsilon_i v_i v_i^\top \right\|_2 = \left\| \sum_{i=1}^n \frac{\varepsilon_i}{\sqrt{f(x_i^\top \phi_*)}} \cdot \left(\sqrt{f(x_i^\top \phi_*)} \boldsymbol{A}_i \right) \right\|_2. \tag{B.59}$$

Note that $\dfrac{\varepsilon_i}{\sqrt{f(x_i^\top \phi_*)}}$ is a standard normal random variable, for all i. We also define a $d \times d$ matrix $\boldsymbol{\Sigma} = \sum_{i=1}^n f(x_i^\top \phi_*) \boldsymbol{A_i^2}$. Then, we have

$$\boldsymbol{\Sigma} = \sum_{i=1}^n f(x_i^\top \phi_*) \left(v_i v_i^\top \right) \left(v_i v_i^\top \right) \tag{B.60}$$

$$= \sum_{i=1}^n f(x_i^\top \phi_*) \left\| v_i \right\|_2^2 v_i v_i^\top. \tag{B.61}$$

We also know

$$\left\| \sum_{i=1}^n \boldsymbol{A}_i \right\|_2 = \left\| \sum_{i=1}^n v_i v_i^\top \right\|_2 \tag{B.62}$$

$$= \left\| \left(\boldsymbol{V}^{-1/2} \boldsymbol{X}^\top \right) \left(\boldsymbol{X} \boldsymbol{V}^{-1/2} \right) \right\|_2 \tag{B.63}$$

$$\leq \left\| \left(\boldsymbol{V}^{-1/2} \boldsymbol{X}^\top \right) \right\|_2 \left\| \left(\boldsymbol{X} \boldsymbol{V}^{-1/2} \right) \right\|_2 \leq 1, \tag{B.64}$$

where (B.64) follows from Lemma 15. Moreover, we know

$$\left\| \boldsymbol{\Sigma} \right\|_2 = \left\| \sum_{i=1}^n f(x_i^\top \phi_*) \left\| v_i \right\|_2^2 v_i v_i^\top \right\|_2 \tag{B.65}$$

$$\leq \left\| d \cdot \sigma_{\max}^2 \sum_{i=1}^{n} v_i v_i^T \right\|_2 \tag{B.66}$$

$$= d \cdot \sigma_{\max}^2 \left\| \sum_{i=1}^{n} \boldsymbol{A}_i \right\| \leq d \cdot \sigma_{\max}^2, \tag{B.67}$$

where (B.66) follows from Lemma 15-16, $f(x_i^\top \phi_*) \leq \sigma_{\max}^2$, and that $v_i v_i^\top$ is positive semi-definite, and the last inequality follows directly from (B.64). By Lemma 17 and the fact that $\varepsilon(x_1), \cdots, \varepsilon(x_n)$ are mutually independent given the contexts $\{x_i\}_{i=1}^n$, we know that

$$\mathbb{P}\Big\{ \lambda_{\max}\Big(\sum_{i=1}^{n} \varepsilon_i \boldsymbol{A}_i \Big) \geq \sqrt{2 \left\| \boldsymbol{\Sigma} \right\|_2 s} \Big\} \leq d \cdot e^{-s}. \tag{B.68}$$

Therefore, by choosing $s = \ln(d/\delta)$ and the fact that $\lambda_{\max}\Big(\sum_{i=1}^{n} \varepsilon_i \boldsymbol{A}_i \Big) = \left\| \sum_{i=1}^{n} \varepsilon_i \boldsymbol{A}_i \right\|_2$, we obtain

$$\mathbb{P}\Big\{ \left\| \sum_{i=1}^{n} \varepsilon_i \boldsymbol{A}_i \right\|_2 \geq \sqrt{2\sigma_{\max}^2 d \ln(\frac{d}{\delta})} \Big\} \leq \delta. \tag{B.69}$$

Finally, by applying Lemma 5 and (B.69) to (B.57), we conclude that for any $n \in \mathbb{N}$, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\left\| \boldsymbol{X}_n^\top \big(\varepsilon \circ \boldsymbol{X}_n (\theta_* - \widehat{\theta}_n) \big) \right\|_{\boldsymbol{V}_n^{-1}} \leq \alpha_n^{(1)}(\delta) \cdot \alpha^{(3)}(\delta). \tag{B.70}$$

□

B.4 Proof of Lemma 9

We first introduce a useful lemma on the norm of the Hadamard product of two matrices.

Lemma 18. *Given any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimension, the following holds:*

$$\left\| \boldsymbol{A} \circ \boldsymbol{B} \right\|_F \leq \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\top) \leq \left\| \boldsymbol{A} \right\|_2 \cdot \left\| \boldsymbol{B} \right\|_2, \tag{B.71}$$

where $\|\cdot\|$ denotes the Frobenius norm. When $\boldsymbol{A}$ and $\boldsymbol{B}$ are vectors, the above degenerates to

$$\|\boldsymbol{A} \circ \boldsymbol{B}\|_2 \leq \|\boldsymbol{A}\|_2 \cdot \|\boldsymbol{B}\|_2 . \tag{B.72}$$

Proof of Lemma 9. To simplify notation, we use $\boldsymbol{X}$ and $\boldsymbol{V}$ as a shorthand for $\boldsymbol{X}_n$ and $\boldsymbol{V}_n$, respectively. Let $\boldsymbol{M}$ be a positive definite matrix. We have

$$\|\boldsymbol{A}v\|_{\boldsymbol{M}} = \left\|\boldsymbol{M}^{1/2}\boldsymbol{A}v\right\|_2 \leq \left\|\boldsymbol{M}^{1/2}\boldsymbol{A}\right\|_2 \cdot \|v\|_2 , \tag{B.73}$$

where the last inequality holds since ℓ_2-norm is sub-multiplicative. Meanwhile, we also observe that

$$\begin{aligned}
&\left(\theta_* - \widehat{\theta}\right)^\top \boldsymbol{X}^\top \boldsymbol{X}\left(\theta_* - \widehat{\theta}\right) && \text{(B.74)}\\
&= \left(\theta_* - \widehat{\theta}\right)^\top \boldsymbol{V}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{V}^{-1/2}\boldsymbol{V}^{1/2}\left(\theta_* - \widehat{\theta}\right) && \text{(B.75)}\\
&= \left\|\left(\theta_* - \widehat{\theta}\right)^\top \boldsymbol{V}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top\right\|_2^2 && \text{(B.76)}\\
&\leq \left\|\left(\theta_* - \widehat{\theta}\right)^\top \boldsymbol{V}^{1/2}\right\|_2^2 \left\|\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top\right\|_2^2 && \text{(B.77)}\\
&\leq \left\|\theta_* - \widehat{\theta}\right\|_{\boldsymbol{V}}^2 . && \text{(B.78)}
\end{aligned}$$

Therefore, we know

$$\begin{aligned}
&\left\|\boldsymbol{X}^\top\left(\boldsymbol{X}\left(\theta_* - \widehat{\theta}\right) \circ \boldsymbol{X}\left(\theta_* - \widehat{\theta}\right)\right)\right\|_{\boldsymbol{V}^{-1}} && \text{(B.79)}\\
&\leq \left\|\boldsymbol{V}^{-1/2}\boldsymbol{X}^\top\right\|_2 \left\|\left(\boldsymbol{X}\left(\theta_* - \widehat{\theta}\right) \circ \boldsymbol{X}\left(\theta_* - \widehat{\theta}\right)\right)\right\|_2 && \text{(B.80)}\\
&\leq 1 \cdot \left\|\boldsymbol{X}\left(\theta_* - \widehat{\theta}\right)\right\|_2^2 && \text{(B.81)}
\end{aligned}$$

$$\leq 1 \cdot \left(\left(\theta_* - \widehat{\theta}\right)^\top \boldsymbol{X}^\top \boldsymbol{X} \left(\theta_* - \widehat{\theta}\right) \right) \tag{B.82}$$

$$\leq \left\| \theta_* - \widehat{\theta} \right\|_{\boldsymbol{V}}^2 \leq (\alpha_n^{(1)}(\delta))^2, \tag{B.83}$$

where (B.81) follows from Lemma 15 and 18, and (B.83) follows from Lemma 5. The proof is complete. □

B.5 Proof of Theorem 2

Recall that $h_\beta(u, v) = \left(\Phi\left(\frac{\beta - u}{\sqrt{f(v)}} \right) \right)^{-1}$. We first need the following lemma about Lipschitz smoothness of the function $h_\beta(u, v)$.

Lemma 19. *The function $h_\beta(u, v)$ defined in (3.18) is (uniformly) Lipschitz smooth on its domain, i.e., there exists a finite $M_h > 0$ (M_h is independent of u, v, and β) such that for any β with $|\beta| \leq B$, for any $u_1, u_2 \in [-1, 1]$ and $v_1, v_2 \in [\sigma_{\min}^2, \sigma_{\max}^2]$,*

$$|\nabla h_\beta(u_1, v_1) - \nabla h_\beta(u_2, v_2)| \leq M_h \left\| \begin{pmatrix} u_1 \\ v_1 \end{pmatrix} - \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} \right\|_2. \tag{B.84}$$

Moreover, we have

$$h_\beta(u_2, v_2) - h_\beta(u_1, v_1) \leq \tag{B.85}$$

$$\begin{pmatrix} u_2 - u_1 \\ v_2 - v_1 \end{pmatrix}^\top \nabla h_\beta(u_1, v_1) + \frac{M_h}{2} \left\| \begin{pmatrix} u_2 - u_1 \\ v_2 - v_1 \end{pmatrix} \right\|_2^2. \tag{B.86}$$

Proof of Lemma 19. First, it is easy to verify that $h_\beta(\cdot, \cdot)$ is twice continuously differentiable on its domain $[-1, 1] \times [\sigma_{\min}^2, \sigma_{\max}^2]$ and therefore is Lipschitz smooth, for some finite positive constant M_h. To show that there exists an M_h that is independent of u, v, β, we need to consider the gradient and Hessian of $h_\beta(\cdot, \cdot)$. Since $h_\beta(u, v)$ is a composite function that involves $\Phi(\cdot)$ and $f(\cdot)$, it is straightforward to write down the first and second derivatives of $h_\beta(u, v)$ with respect to u and v, which depend on $\Phi(\cdot)$, $\Phi'(\cdot)$, $\Phi''(\cdot)$, $f(\cdot)$, $f'(\cdot)$, and $f''(\cdot)$. Given the facts that for all the u, v and β in the domain of interest, we have $\Phi(\frac{\beta - u}{v}) \in [\Phi(\frac{-B-1}{\sigma_{\min}^2}), 1]$, $\Phi'(\frac{\beta - u}{v}) \in (0, \frac{1}{\sqrt{2\pi}})$,

$|\Phi''(\frac{\beta-u}{v})| \leq \frac{B+1}{\sigma_{\min}\sqrt{2\pi}}$, and that $f(\cdot), f'(\cdot), f''(\cdot)$ are all bounded, it is easy to verify that such an M_h indeed exists by substituting the above conditions into the first and second derivatives of $h_\beta(u,v)$ with respect to u and v. Moreover, by Lemma 3.4 in [88], we know that (B.86) indeed holds. □

Proof of Theorem 2. Define

$$q_u := \sup_{u_0 \in (-1,1)} |\frac{\partial h_\beta}{\partial u}|\Big|_{u=u_0}, \tag{B.87}$$

$$q_v := \sup_{v_0 \in (\sigma_{\min}^2, \sigma_{\max}^2)} |\frac{\partial h_\beta}{\partial v}|\Big|_{v=v_0}. \tag{B.88}$$

By the discussion in the proof of Lemma 19, we know that q_u and q_v are both positive real numbers. By substituting $u_1 = \theta_1^\top x$, $u_2 = \theta_2^\top x$, $v_1 = f(\phi_1^\top x)$, and $v_2 = f(\phi_2^\top x)$ into (B.86), we have

$$h_\beta\big(\theta_2^\top x, \phi_2^\top x\big) - h_\beta\big(\theta_1^\top x, \phi_1^\top x\big) \tag{B.89}$$

$$\leq \begin{pmatrix} (\theta_2-\theta_1)^\top x \\ f(\phi_2^\top x) - f(\phi_1^\top x) \end{pmatrix}^\top \nabla h_\beta(\theta_1^\top x, f(\phi_1^\top x)) \tag{B.90}$$

$$+ \frac{M_h}{2} \left\| \begin{pmatrix} (\theta_2-\theta_1)^\top x \\ f(\phi_2^\top x) - f(\phi_1^\top x) \end{pmatrix} \right\|_2^2 \tag{B.91}$$

$$\leq \big(q_u \left\|\theta_2-\theta_1\right\|_{\boldsymbol{M}} \cdot \left\|x\right\|_{\boldsymbol{M}^{-1}} \tag{B.92}$$

$$+ q_v M_f \left\|\phi_2-\phi_1\right\|_{\boldsymbol{M}} \cdot \left\|x\right\|_{\boldsymbol{M}^{-1}} \big) \tag{B.93}$$

$$+ \frac{M_h}{2} \big(\left\|\theta_2-\theta_1\right\|_{\boldsymbol{M}}^2 + M_f^2 \left\|\phi_2-\phi_1\right\|_{\boldsymbol{M}}^2 \big) \cdot \left\|x\right\|_{\boldsymbol{M}^{-1}} \tag{B.94}$$

$$\leq (q_u + M_h) \left\|\theta_2-\theta_1\right\|_{\boldsymbol{M}} \cdot \left\|x\right\|_{\boldsymbol{M}^{-1}} \tag{B.95}$$

$$+ M_f(q_v + M_h M_f L) \left\|\phi_2-\phi_1\right\|_{\boldsymbol{M}} \cdot \left\|x\right\|_{\boldsymbol{M}^{-1}}, \tag{B.96}$$

where (B.93)-(B.94) follow from the Cauchy-Schwarz inequality and the fact that $f(\cdot)$ is Lipschitz continuous, and (B.95)-(B.96) follow from the facts that $\left\|x\right\|_2 \leq 1$, $\left\|\theta_2-\theta_1\right\|_2 \leq 2$, and

$\|\phi_2 - \phi_1\|_2 \leq 2L$. By letting $C_3 = q_u + M_h$ and $C_4 = M_f(q_v + M_h M_f L)$, we conclude (3.34)-(3.35) indeed holds with C_3 and C_4 being independent of $\theta_1, \theta_2, \phi_1, \phi_2$, and β. □

B.6 Proof of Lemma 10

Proof. By Theorem 2 and (3.20), we know

$$Q^{\text{HR}}_{t+1}(x) - h_{\beta_{t+1}}(\theta_*^\top x, \phi_*^\top x) \tag{B.97}$$

$$= h_{\beta_{t+1}}(\widehat{\theta}_t^\top x, \widehat{\phi}_t^\top x) + \xi_t(\delta) \|x\|_{\boldsymbol{V}_t^{-1}} - h_{\beta_{t+1}}(\theta_*^\top x, \phi_*^\top x) \tag{B.98}$$

$$\leq 2\xi_t(\delta) \|x\|_{\boldsymbol{V}_t^{-1}}. \tag{B.99}$$

Similarly, by switching the roles of $\theta_*^\top, \phi_*^\top$ and $\widehat{\theta}_t^\top, \widehat{\phi}_t^\top$ in (B.98), we have

$$Q^{\text{HR}}_{t+1}(x) - h_{\beta_{t+1}}(\theta_*^\top x, \phi_*^\top x) \geq 0. \tag{B.100}$$

□

B.7 Proof of Theorem 3

Proof. For each user t, let $\pi_t^{\text{HR}} = \{x_{t,1}, x_{t,2}, \cdots\}$ denote the action sequence under the HR-UCB policy. Under HR-UCB, $\widehat{\theta}_t$ and $\widehat{\phi}_t$ are updated only after the departure of each user. This fact implies that $x_{t,i} = x_{t,j}$, for all i, j. Therefore, we can use x_t to denote the action chosen by HR-UCB for the user t, to simplify notation. Let $\overline{R}_t^{\text{HR}}$ denote the expected lifetime of user t under HR-UCB. Similar to (3.9), we have

$$\overline{R}_t^{\text{HR}} = \Big(\Phi\Big(\frac{\beta_t - \theta_*^\top x_t}{\sqrt{f(\phi_*^\top x_t)}}\Big)\Big)^{-1} = h_{\beta_t}(\theta_*^\top x_t, \phi_*^\top x_t). \tag{B.101}$$

Recall that π^{oracle} and x_t^* denote the oracle policy and the context of the action of the oracle policy for user t, respectively. We compute the pseudo regret of HR-UCB as

$$\text{Regret}_T = \sum_{t=1}^{T} \overline{R}_t^* - \overline{R}_t^{\text{HR}} \tag{B.102}$$

$$= \sum_{t=1}^{T} h_{\beta_t}\big(\theta_*^\top x_t^*, \phi_*^\top x_t^*\big) - h_{\beta_t}\big(\theta_*^\top x_t, \phi_*^\top x_t\big). \tag{B.103}$$

To simplify notation, we use w_t as a shorthand for $h_{\beta_t}\big(\theta_*^\top x_t^*, \phi_*^\top x_t^*\big) - h_{\beta_t}\big(\theta_*^\top x_t, \phi_*^\top x_t\big)$. Given any $\delta > 0$, define an event E_δ in which (3.21) and (3.22) hold under the given δ, for all $t \in \mathbb{N}$. By Lemma 5 and Theorem 1, we know that the event E_δ occurs with probability at least $1 - 3\delta$. Therefore, with probability at least $1 - 3\delta$, for all $t \in \mathbb{N}$,

$$w_t \le Q_t^{\text{HR}}(x_t^*) - h_{\beta_t}\big(\theta_*^\top x_t, \phi_*^\top x_t\big) \tag{B.104}$$

$$\le Q_t^{\text{HR}}(x_t) - h_{\beta_t}\big(\theta_*^\top x_t, \phi_*^\top x_t\big) \tag{B.105}$$

$$= h_{\beta_t}\big(\theta_*^\top x_t, \phi_*^\top x_t\big) + \xi_{t-1}(\delta)\, \|x_t\|_{\boldsymbol{V}_{t-1}^{-1}} \tag{B.106}$$

$$\quad - h_{\beta_t}\big(\theta_*^\top x_t, \phi_*^\top x_t\big) \tag{B.107}$$

$$\le 2\xi_{t-1}(\delta) \cdot \|x_t\|_{\boldsymbol{V}_{t-1}^{-1}}, \tag{B.108}$$

where (B.104) and (B.106) follow directly from the definition of the UCB index, (B.105) follows from the design of HR-UCB algorithm, and (B.108) is a direct result under the event E_δ. Now, we are ready to conclude that with probability at least $1 - 3\delta$, we have

$$\text{Regret}_T = \sum_{t=1}^{T} w_t \le \sqrt{T \sum_{t=1}^{T} w_t^2} \tag{B.109}$$

$$\le \sqrt{4\xi_T^2(\delta) T \sum_{t=1}^{T} \min\{\|x_t\|^2_{\boldsymbol{V}_{t-1}^{-1}}, 1\}} \tag{B.110}$$

$$\le \sqrt{8\xi_T^2(\delta) T \cdot d \log\Big(\frac{\mathcal{S}(T) + \lambda d}{\lambda d}\Big)}, \tag{B.111}$$

where (B.109) follows from the Cauchy-Schwarz inequality, (B.110) follows from the fact that $\xi_t(\delta)$ is an increasing function in t, and (B.111) follows from Lemma 10 and 11 in [27] and the fact that $\boldsymbol{V}_t = \lambda \boldsymbol{I}_d + \boldsymbol{X}_t^\top \boldsymbol{X}_t = \lambda \boldsymbol{I}_d + \sum_{i=1}^t x_i x_i^\top$. By substituting $\xi_T(\delta)$ into (B.111) and using the fact that $\mathcal{S}(T) \leq \Gamma(T)$, we know

$$\text{Regret}_T = O\bigg(\sqrt{T \log \Gamma(T) \cdot \Big(\log\big(\Gamma(T)\big) + \log(\frac{1}{\delta})\Big)^2}\bigg). \tag{B.112}$$

By choosing $\Gamma(T) = KT$ for some constant $K > 0$, we thereby conclude that

$$\text{Regret}_T = O\bigg(\sqrt{T \log T \cdot \Big(\log T + \log(\frac{1}{\delta})\Big)^2}\bigg). \tag{B.113}$$

The proof is complete. □