Fundamental Limits in Stochastic Bandits

PO-AN WANG

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 11th October 2024, at 10:00 in D37, Lindstedtsvägen 9, Stockholm.

Doctoral Thesis in Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden 2024

ISBN 978-91-8106-033-1
TRITA-EECS-AVL-2024:64

Printed by: Universitetsservice US-AB, Sweden 2024

Abstract

This thesis contributes to the field of stochastic bandits by exploring the fundamental limits (information-theoretic lower bounds) of three prevalent objects in various reinforcement learning applications, through a collection of five distinct papers. Each paper presents a novel perspective under a specific structure and scenario. The first paper investigates regret minimization in decentralized multi-agent multi-armed bandits. The second and third papers delve into pure exploration with fixed confidence in a broad spectrum of structured bandits. The last two papers focus on offering new insights into the best arm identification with a fixed budget.

In the first paper, two popular scenarios in a decentralized multi-agent setting are addressed, one involving collision and the other not. In each of them, we propose an instance-specific optimal algorithm. Interestingly, our results show that the fundamental limits match the ones in the centralized analogue. The second paper introduces a simple but versatile algorithm, Frank-Wolfe Sampling, which achieves instance-specific optimality across a wide collection of pure explorations in structured bandits. Meanwhile, the numerical results and current studies demonstrate the strong numerical performance of our algorithm in various pure exploration problems. However, Frank-Wolfe Sampling is not computationally efficient when the number of arms is extremely large. To address this issue, the third paper introduces Perturbed Frank-Wolfe Sampling, which can be implemented in polynomial time while maintaining the instance-specific minimal sample complexity in combinatorial semi-bandits.

Unlike the sample complexity or regret minimization discussed above, characterizing the fundamental limit of the error probability in best arm identification with a fixed budget remains a challenge. The fourth paper addresses this challenge in two-armed bandits, introducing a new class of algorithms, stable algorithms, which encompass a broad range of reasonable algorithms. We demonstrate that no consistent and stable algorithm surpasses the algorithm that samples each arm evenly, answering the open problems formulated in prior work. In general multi-armed bandits, the final paper in this thesis presents, to our knowledge, the first large deviation theorem for the generic adaptive algorithm. Based on this, we provide the exact analysis of the celebrated algorithm, Successive Rejects. Furthermore, this new large deviation technique allows us to devise and analyze a new adaptive algorithm, which is the current state-of-the-art to the best of our knowledge. This thesis provides new insight for deriving fundamental limits in various online stochastic learning problems. This understanding guides us to develop more efficient algorithms and systems.

Sammanfattning

Denna avhandling bidrar till området för Förstärkningsinlärning (RL) genom att utforska de grundläggande gränserna (informationsteoretiska nedre gränser) för tre vanliga objekt i olika förstärkningsinlärningsapplikationer, genom en samling av fem distinkta uppsatser. Varje uppsats presenterar ett nytt perspektiv under en specifik struktur och scenario. Den första uppsatsen undersöker ångerminimering i decentraliserade multi-agent multi-armed banditer. Den andra och tredje uppsatsen dyker in i ren utforskning med fast förtroende i ett brett spektrum av strukturerade banditer. De två sista uppsatserna fokuserar på att erbjuda nya insikter i identifieringen av den bästa armen med en fast budget.

I den första uppsatsen behandlas två populära scenarier i en decentraliserad multi-agent inställning, en som involverar kollision och den andra inte. I vardera av dem föreslår vi en instansspecifik optimal algoritm. Intressant nog visar våra resultat att de grundläggande gränserna matchar de som finns i den centraliserade analogin. Den andra uppsatsen introducerar en enkel men mångsidig algoritm, Frank-Wolfe Sampling, som uppnår instansspecifik optimalitet över en bred samling av rena utforskningar i strukturerade banditer. Samtidigt demonstrerar de numeriska resultaten och aktuella studier den starka numeriska prestandan för vår algoritm i olika rena utforskningsproblem. Dock är Frank-Wolfe Sampling inte beräkningsmässigt effektiv när antalet armar är extremt stort. För att lösa detta problem introducerar den tredje uppsatsen Perturbed Frank-Wolfe Sampling, vilket kan implementeras i polynomisk tid samtidigt som den instansspecifika minimala provkomplexiteten bibehålls i kombinatoriska semi-banditer.

Till skillnad från provkomplexitet eller ångerminimering som diskuterats ovan, förblir karaktäriseringen av den grundläggande gränsen för felprocenten vid identifiering av den bästa armen med en fast budget en utmaning. Den fjärde uppsatsen tar upp denna utmaning i två-armede banditer, genom att introducera en ny klass av algoritmer, stabila algoritmer, som omfattar ett brett spektrum av rimliga algoritmer. Vi demonstrerar att ingen konsekvent och stabil algoritm överträffar algoritmen som provtar varje arm jämnt, vilket svarar på de öppna problem som formulerats i tidigare arbete. I allmänna multi-armede banditer presenterar den sista uppsatsen i denna avhandling, till vår kännedom, den första stora avvikelseteoremen för den generiska adaptiva algoritmen. Baserat på detta ger vi den exakta analysen av den berömda algoritmen, Successive Rejects. Dessutom låter denna nya stora avvikelseteknik oss att utforma och analysera en ny adaptiv algoritm, vilket är den nuvarande state-of-the-art till bästa av vår kunskap. Denna avhandling ger ny insikt för att härleda grundläggande gränser i olika online stokastiska inlärningsproblem. Denna förståelse vägleder oss att utveckla mer effektiva algoritmer och system.

List of Papers

This thesis is founded on the research from five distinct papers.

Paper I

Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo, "Optimal Algorithms for Multiplayer Multi-Armed Bandits," In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Paper II

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere, "Fast Pure Exploration via Frank-Wolfe" In *The 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Paper III

Ruo-Chun Tzeng, Po-An Wang, Alexandre Proutiere, and Chi-Jen Lu, "Closing the Computational-Statistical Gap in Best Arm Identification for Combinatorial Semi-bandits" In *The 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Paper IV

Po-An Wang, Kaito Ariu, and Alexandre Proutiere, "On Universally Optimal Algorithms for A/B Testing" In *The 41st International Conference on Machine Learning (ICML)*, 2024

Paper V

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere, "Best Arm Identification with Fixed Budget: A Large Deviation Perspective" In *The 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

The following publications and presentations are not included in this thesis.

Conference Papers

1. Po-An Wang, and Chi-Jen Lu, "Tensor Decomposition via Simultaneous Power Iteration," In *the 34th International Conference on Machine Learning (ICML)*, 2017.

2. Yi-Shan Wu, Po-An Wang, and Chi-Jen Lu, "Lifelong Optimization with low Regret", In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

3. Ruo-Chun Tzeng, Po-An Wang, Florian Adriaens, Aristides Gionis, Chi-Jen Lu, "Improved analysis of randomized SVD for top-eigenvector approximation, " In *The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Preprints

1. Alexandre Proutiere, and Po-An Wang "An Optimal Algorithm for Multi-player Multi-Armed Bandits ," *arXiv preprint arXiv:1909.13079*, 2019.

2. Po-An Wang, Kaito Ariu, and Alexandre Proutiere, "On Uniformly Optimal Algorithms for Best Arm Identification in Two-Armed Bandits with Fixed Budget," *arXiv preprint arXiv:2308.12000*, 2023.

Acknowledgement

The journey of pursuing a PhD has been one of the most beautiful and gratifying experiences of my life. Despite the numerous frustrations and challenges, the support and encouragement from many individuals have been invaluable, and without them, this thesis would not have been possible.

First and foremost, I would like to express my deepest gratitude to my supervisor, Alexandre Proutiere. His guidance, theoretical rigor, and expertise have been the bedrock of this thesis. The unwavering support and insightful advice from him have steered me through the complexities of my research and have significantly contributed to my academic and personal growth.

Several excellent researchers have inspired and assisted me along this journey. I am particularly grateful to Aristides Gionis and Chi-Jen Lu for their valuable insights, collaboration, and encouragement. Their expertise and willingness to share knowledge have greatly enriched my research experience.

One of the most wonderful and life-changing decisions I have ever made was joining KTH. In our group, I have had the privilege of working with intelligent and kind colleagues, including Yassir Jedra, Damianos Tranos, Kaito Ariu, Filippo Vannella, Stefan Stojanovic, Daniele Foffano, Frédéric Zheng, Bastien Dubail, and William Réveillard. Whenever I needed a discussion, they were always there to provide support, share ideas, and offer constructive feedback. Their camaraderie and collaboration have made this journey both intellectually rewarding and enjoyable. Moreover, I would like to extend my heartfelt thanks to my warm and supportive colleagues, Hampei Sasahara, Sarit Khirirat, Liam Taghavian, Robert Marczuk Bereza-Jarocinski, Erik Berglund, Siyuan Liu, Ying Wang, Jiabao He, Elis Stefansson, Nana Wang, Braghadeesh Lakshminarayanan, Xiao Chen, Kun Cao, Sijing Tu, Honglian Wang, Guo-Jhen Wu, Linda Tenhu, Nicola Bastianello, Javad Parsa, Mani Hemanth Dhullipalla, Peihu Duan, Dongjae Lee, Takuya Iwaki, Jiaojiao Zhang, Yu Wang, Yuchao Lee, Mohit Daga, and Amir Mohammad Karimi Mamaghan. Their friendship and support have been invaluable throughout this journey, and I am grateful for the many moments of laughter and encouragement we have shared.

My special thanks go to Kaito Ariu, who invited me to do an internship at CyberAgent in Tokyo. This experience provided me with invaluable insights and opportunities in the industry that have significantly contributed to my personal and professional development. During my time there, I had the chance to work with amazing and friendly people, including Yuma Fujimoto, Jinna Yuu, Kenshi Abe, Tetsuro Morimura, Mitsuki Sakamoto, Qiqi Gao, Chao Qin, and Shun Kitamura. Their support and collaboration were truly enriching.

I cherish my Taiwanese friends, whose companionship and support have been a constant source of strength and joy throughout this journey. Their kindness, encouragement, and the countless moments we shared have made this experience truly special. Thank you for being such an important part of my life and for always being there for me.

Finally, I would like to thank my family for their constant support. Their warmth and encouragement have been a beacon of light, lifting my spirits during challenging times. I am also deeply grateful to my partner, Ruo-Chun Tzeng, for her companionship on this journey and her invaluable assistance in my research. With her presence and encouragement, all the difficulties seemed like a piece of cake.

Po-An Wang,
September 2023.

Outline

Chapter 1

Introduction

Drawing parallels from the way living beings adapt and learn from their experiences in real life, Reinforcement Learning (RL) algorithms consistently improve their performance by methodically collecting and analyzing data from their environment [59, 70]. Much like the versatility of human capabilities, RL has been extensively utilized across a diverse range of fields such as strategic game playing (like AlphaGO [66, 67]), autonomous navigation (like self-driving cars [54]), robotics (for tasks like object manipulation [46]), resource management in computer systems [50], finance (for automated trading systems [47]), healthcare (for personalized treatment strategies [24, 83]), and even in the development of large language models [13, 56].

Despite the successful application of RL in these various tasks, its fundamental limitations, stemming from the complex interaction between the agent and its environment, continue to pose challenges for researchers. Understanding these fundamental limits not only aids in improving the efficiency of RL by guiding algorithm design but also helps in setting realistic expectations for its application.

To further explore these challenges, this thesis uses the multi-armed bandits (MAB) problem as a stepping stone. The heart of RL is the essential concept of learning from trial and error, which is encapsulated in the MAB problem. MAB algorithms find frequent application in areas requiring decision-making under uncertainty, such as recommendation systems [1, 65], online advertising [84], and medical trials [75].

This thesis narrows its focus to a specific type of MAB problem: stochastic bandits. In this scenario, the rewards are random variables that can change over time. The inherent uncertainty and variability in stochastic bandits present unique challenges and opportunities for RL algorithms, offering a fertile ground to explore the fundamental limits of RL and to develop strategies to overcome

these limits [30].

A significant challenge in stochastic bandits is the exploration-exploitation trade-off. An RL agent is faced with the decision to either exploit known sources of reward or to explore new options that might yield higher rewards in the future. Effectively balancing this trade-off is crucial for the performance of RL algorithms.

Despite extensive research in stochastic bandits, there remain open questions about the fundamental limits of what can be achieved. What is the best possible performance that can be achieved by any RL algorithm in a stochastic bandit setting? How does the complexity of the environment affect these limits? How can we design algorithms that approach these limits?

In this thesis, we aim to answer these questions. We conduct a comprehensive analysis of various RL algorithms in stochastic bandit settings, identifying their strengths and limitations. We also propose algorithms that push the boundaries of what is currently achievable. Through our research, we aspire to deepen the understanding of the fundamental limits in stochastic bandits.

We believe that our findings will not only guide the design of more efficient RL algorithms but also help in setting more realistic expectations for the application of RL. Ultimately, we hope that our research will contribute to the effective and responsible application of RL across various fields.

1.1 Stochastic Bandits and Three Objectives

In the MAB problem, a set of arms exists, each assigned an index ranging from 1 to K. A learner is required to sequentially pull an arm from the set $[K] \coloneqq \{1, \dots, K\}$ and observe a random reward generated according to the corresponding distribution. More specifically, for each $k \in [K]$, ν_k denotes the reward distribution on arm k, with an average of μ_k. These averages, represented by the vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$, is initially unknown to the learner. For convenience, we assume that the arm yielding the highest reward mean is unique and denoted by $1(\boldsymbol{\mu})$. The learner selects $A_t \in [K]$ to be explored in round t based on past observations, yielding the observed reward X_t. The arm A_t selected in round t is measurable with respect to $\mathcal{F}_t$, where $\mathcal{F}_t$ represents the σ-algebra generated by the set of random variables $\{A_t, X_1, \dots, A_{t-1}, X_{t-1}\}$. Subsequently, we will discuss the three objectives this thesis aims to optimize.

Regret Minimization.
Regret is defined as the difference between the reward a learner could have received if the optimal arm $1(\boldsymbol{\mu})$ was known, and the reward she actually received. The common objective is to minimize this expected regret. This is represented

as:

$$R^{\pi}(T) := T\mu_{1(\boldsymbol{\mu})} - \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} X_t\right],$$

where $\mathbb{E}_{\boldsymbol{\mu}}$ denotes the expectation under parameter $\boldsymbol{\mu}$ (likewise, $\mathbb{P}_{\boldsymbol{\mu}}$ represents the probability under $\boldsymbol{\mu}$), and π symbolizes the implemented strategy. To achieve this, the learner must strike a balance between exploration (trying new arms) and exploitation (pulling the arm with the highest expected reward), due to the inherent uncertainty of the environment. The fundamental limit of this problem, also known as the information-theoretic lower bound, was proposed in [43]. This limit essentially indicates the necessary exploration cost for each arm. Numerous algorithms that match this lower bound, such as KL-UCB [28], Thompson Sampling [40, 73], and IMED [32], have demonstrated strong and robust performance across various scenarios.

Fixed Confidence Best Arm Identification.
In some situations, the primary focus is to answer specific questions about the underlying distributions. These are known as *pure exploration* problems. Among these, Best Arm Identification (BAI) is arguably the most well-known. In fixed confidence best arm identification, a learner aims to estimate the best arm with an accuracy higher than a certain $\delta \in (0, 1)$, while minimizing the expected number of pulls. More precisely, an algorithm consists of a sampling rule, dictating the arm $A_t \in \mathcal{F}_t$ to be explored in round t, a stopping rule determining the round τ to stop upon gathering enough observations, and the decision rule $\hat{\imath} \in \mathcal{F}_\tau$. The performance is measured by $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ subject to $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})] \leq \delta$. The lower bound for this problem was shown in [29] and first matched by the Track-and-Stop (TaS) algorithm proposed in the same work.

Fixed Budget Best Arm Identification.
As pointed out in [60], this problem can be seen as the dual problem of fixed confidence. Instead of minimizing the number of pulls to reach a given accuracy, the learner aims to maximize the accuracy using a given budget for the number of pulls. As the stopping rule defaults to a given budget T, an algorithm consists of only two rules: the sampling rule, $A_t \in \mathcal{F}_t$ and the decision rule $\hat{\imath} \in \mathcal{F}_T$. The performance of an algorithm is measured by the error probability at the end, namely $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})]$. Despite recent research efforts, this problem remains largely open in contrast to the above two objectives. To the best of our knowledge, its fundamental limit is characterized only in two-armed bandits, which is the main focus of the fourth paper in this thesis [76].

1.2 Challenges

In this section, we briefly discuss the inherent challenges associated with developing algorithms designed to achieve the fundamental limit. These challenges represent more than just technical hurdles; they serve as critical junctures that bridge the gap between theoretical construction and practical applications.

1.2.1 Limited Communication in the Cooperative Multi-Agent Setting

In a cooperative multi-agent setting, the arm chosen by one agent can significantly influence the rewards of others. This interconnected dynamic underscores the critical role of effective communication among agents. Such communication is not only essential for coordinating their joint behavior but also for optimizing the collective objective. However, in real-world scenarios, the cost of communication can be steep, particularly in terms of the total bits of messages exchanged among agents. Beyond the sheer cost, it may also involve protocol design and the necessity for efficient data encoding. When there is no communication limit, also known as a centralized setting, the fundamental limit has been well studied in [5]. The first paper of this thesis [77], aims to answer the fundamental limits in two types of decentralized settings.

1.2.2 Information Gain from the Structure

The structure of the parameter set plays a vital role in encoding the side information among the reward means. Effectively leveraging this structure is a crucial step towards achieving scalability in real applications. Consider a recommendation system, for instance. In such a system, the parameter set can be used to encode user preferences or item characteristics, providing valuable information that can improve the quality of recommendations, leading to increased user satisfaction and system efficiency. Furthermore, the scalability of the system can be enhanced by optimizing the use of this side information, allowing the system to handle a larger number of users and items without sacrificing performance. This is particularly important as recommendation systems nowadays are required to process vast amounts of data and cater to the needs of millions of users. The second and third papers in this thesis [74, 78] aim to characterize the information gained from structure.

1.2.3 Flexibility Across a Wide Class of Problems

Needless to say, an ideal method would be one that can adapt to various tasks. An optimal algorithm would not only be theoretically sound but also versatile, capable of adjusting its functionality in response to different problem sets and

environmental conditions. This adaptability is a cornerstone of practical efficiency, enabling the algorithm to maintain optimal performance even when faced with unstudied circumstances. Previous works [14, 23] achieve minimal regret in generic structured bandits, and [55] further extends it to structured ergodic Markov decision processes. The second paper of this thesis attains the minimal sample complexity in generic structured bandits [78].

1.2.4 Computational-Statistical Gap

When dealing with an extremely large set of arms, most statistically efficient algorithms suffer from computational intractability. This is primarily due to these algorithms being designed to explore each arm to estimate its value. However, if the set size increases significantly, the computational resources required to explore and evaluate each arm also increase exponentially. This leads to a situation where the algorithm cannot complete its task within a reasonable time or without excessive resource consumption. Such a computational-statistical gap in general RL is discussed in [36]. In the third paper of this thesis [74], we study the fundamental limit under the computational constraint that the algorithm can be implemented in polynomial time.

1.2.5 Large Deviation Principle for Adaptive Algorithms

In the fixed budget best arm identification problem, the error probability of an algorithm will decay exponentially. While the algorithm is a static sampling whose allocation is fixed, the decay rate can be explicitly determined via classical large deviation theory [31]. However, in the case of an adaptive algorithm, where the sampling rule adapts to the observation, the decay rate becomes far more challenging due to the intriguing dependence between the observations. The fifth paper of this thesis aims at establishing the bound of error rate for generic algorithms [79]. Moreover, as pointed out in the recent study [6, 19], it is impossible to characterize the minimal instance-specific error rate within an algorithm class that includes all static algorithms. The fourth paper of this thesis, therefore, attempts to define a reasonable and wide algorithm class, so that the minimal error rate within that class can be attained by a single algorithm [76].

1.3 Overview of the Thesis

The purpose of this thesis is to investigate the fundamental limits (information-theoretic lower bounds) of three common objectives under various RL applications. This research is situated in the broader context of stochastic learning in RL and seeks to contribute to our understanding of instance-specific optimal algorithms. This thesis consists of five papers, each of which addresses certain challenges mentioned above. The first paper studies regret minimization in decentralized multi-agent MAB. The second and third papers investigate pure ex-

ploration with fixed confidence in a wide class of structured bandits. The final two papers concentrate on providing new perspectives in best arm identification with a fixed budget. In the following, we introduce them in more detail.

In the first paper, two settings are studied in decentralized multi-agent MAB, one with collision and the other without. In each of them, we present an instance-specific optimal algorithm that outperforms the state-of-the-art algorithms [8,51]. Surprisingly, these two algorithms perform as well as the optimal algorithms in centralized settings while maintaining a finite number of communication costs (in terms of the number of collisions in the first setting or bits in the second setting) in expectation.

The second paper introduces a simple algorithm, Frank-Wolfe Sampling (FWS), that attains instance-specific optimality in a wide range of pure explorations in structured bandits. In our experiments, together with the current works [61,82], FWS demonstrates strong numerical performance in various types of pure exploration problems. However, as pointed out in [71], FWS is not computationally efficient when the number of arms is overly large. Inspired by this challenge, the third paper devises Perturbed Frank-Wolfe Sampling (P-FWS), which runs in polynomial time while maintaining the instance-specific minimal sample complexity in combinatorial semi-bandits. Additionally, P-FWS is the first optimal algorithm proven to have polynomial sample complexity in the moderate confidence regime.

In contrast to the sample complexity or regret minimization discussed above, characterizing the fundamental limit of the error probability in best arm identification with a fixed budget is widely open. The fourth paper investigates this difficulty in two-armed bandits. We present a novel definition of an algorithm class, stable algorithm, which includes a wide class of reasonable algorithms. We show that no consistent and stable algorithm outperforms uniform sampling, answering the open problems formulated in [60]. The fifth paper in this thesis presents, to the best of our knowledge, the first large deviation theorem for a generic adaptive algorithm. Based on it, we improve the error probability upper bound of the existing algorithm, Successive Rejects, and this bound is shown to be tight in the fourth paper. Moreover, the fourth paper presents the first truly adaptive algorithm, Continuous Rejects, which is the state-of-the-art algorithm in both theoretical and numerical performance.

Chapter 2

Background: The Lower Bounds

In this chapter, we delve into the intrinsic limitations of stochastic bandits by focusing on the instance-specific lower bounds of the objectives introduced in 1.1. Unlike worst-case lower bounds, which consider only the most challenging instance in the entire problem set, the lower bounds we examine are instance-specific. This means they depend on the specific instance of the problem at hand. The advantage of such instance-specific lower bounds lies in their ability to characterize a more precise understanding of the difficulty associated with a particular problem instance.

All the lower bounds can be derived using the change-of-measure argument [30, 39, 43]. We present a standard one in Lemma 1. For simplicity, throughout this chapter we assume ν_k is a Bernoulli distribution and the mean, μ_k, belongs to $(0, 1)$ for all $k \in [K]$. In such a scenario, the Kullback-Leibler (KL) divergence between ν_k and ν_m is

$$d(\mu_k, \mu_m) = \mu_k \ln \frac{\mu_k}{\mu_m} + (1 - \mu_k) \ln \frac{1 - \mu_k}{1 - \mu_m}.$$

As in Section 1.1, the best arm is assumed to be unique and denoted by $1(\boldsymbol{\mu})$. We introduce

$$\Lambda := \{\boldsymbol{\mu} \in (0, 1)^K : \exists k \in [K] \text{ such that } \mu_k > \mu_m, \forall m \neq k\}$$

as the set of all possible parameters.

Lemma 1 (Lemma 19 in [39])**.** *Let σ be any almost surely stopping time with respect to $\mathcal{F}_t$. For every $\mathcal{F}_\sigma$-measurable random variable Z bounded in $[0, 1]$, and $\boldsymbol{\mu}, \boldsymbol{\lambda} \in \Lambda$,*

$$\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_k(\sigma)] d(\mu_k, \lambda_k) \geq d(\mathbb{E}_{\boldsymbol{\mu}}[Z], \mathbb{E}_{\boldsymbol{\lambda}}[Z]),$$

where $N_k(t)$ denotes the number of times arm k is pulled up to round t.

2.1 Regret Minimization

In this section, we present the instance-specific lower bound from [10]. The proof primarily follows the one in [30]. For regret minimization, an effective strategy is expected to pull a suboptimal arm significantly less than the total number of pulls. In Definition 1, we introduce a class of strategies, referred to as *uniformly fast convergent*, that adhere to this expectation.

Definition 1. *A strategy π is uniformly fast convergent on Λ if for all $\boldsymbol{\mu} \in \Lambda$, $k \neq 1(\boldsymbol{\mu})$, and $\alpha \in (0,1)$, it satisfies that $\mathbb{E}_{\boldsymbol{\mu}}[N_k(T)] = o(T^\alpha)$.*

Theorem 1. *If π is a uniformly fast convergent strategy, then for any $\boldsymbol{\mu} \in \Lambda$, one has*

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\ln T} \geq \sum_{k\neq 1} \frac{\mu_{1(\boldsymbol{\mu})} - \mu_k}{d(\mu_k, \mu_{1(\boldsymbol{\mu})})}.$$

Proof. Recall that

$$R^\pi(T) = \sum_{k\neq 1(\boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\mu}}[N_k(T)](\mu_{1(\boldsymbol{\mu})} - \mu_k).$$

The lower bound is directly obtained once we show

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]}{\ln T} \geq \frac{1}{d(\mu_k, \mu_{1(\boldsymbol{\mu})})}, \forall k \neq 1(\boldsymbol{\mu}) \tag{2.1}$$

Proof of (2.1). Let $k \neq 1(\boldsymbol{\mu})$ be fixed. We choose an alternative $\boldsymbol{\lambda} \in \lambda$ such that $\lambda_m = \mu_m, \forall m \neq k$ and $\lambda_k \in (\mu_{1(\boldsymbol{\mu})}, 1)$. Applying Lemma 1 with $\sigma = T$ and $Z = N_k(T)/T$ yields that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]d(\mu_k, \lambda_k) &\geq d(\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]}{T}, \frac{\mathbb{E}_{\boldsymbol{\lambda}}[N_k(T)]}{T}) \\
&\geq (1 - \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]}{T}) \ln \frac{T}{T - \mathbb{E}_{\boldsymbol{\lambda}}[N_k(T)]} - \ln 2,
\end{aligned} \tag{2.2}$$

where the last inequality is simply due to the fact that $\forall p, q \in (0,1)$

$$d(p,q) = \underbrace{p \ln \frac{1}{q}}_{\geq 0} + (1-p)\frac{1}{1-q} + \underbrace{(p \ln p + (1-p)\ln(1-p))}_{\geq -\ln 2}$$

Since π is a uniformly fast convergent strategy, we deduce that $\mathbb{E}_{\boldsymbol{\mu}}[N_k(T)] = o(T^\alpha)$ for any $\alpha \in (0,1)$. Moreover, using the definition of uniformly fast convergent strategy again, we get

$$T - \mathbb{E}_{\boldsymbol{\lambda}}[N_k(T)] = \sum_{m \neq k} \mathbb{E}_{\boldsymbol{\lambda}}[N_m(T)] = o(T^\alpha),\ \forall \alpha \in (0,1).$$

As a consequence, (2.1) is lower bounded by $(1-\alpha)\ln T$. We hence obtain

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]}{\ln T} \geq \frac{1-\alpha}{d(\mu_k, \lambda_k)}.$$

The proof is then completed as λ_k and α can arbitrarily approach $\mu_{1(\boldsymbol{\mu})}$ and 0 respectively. □

2.2 Fixed Confidence Best Arm Identification

Similar to Section 2.1, we narrow our focus to a class of strategies, as described in Definition 2. In the context of fixed confidence best arm identification, there is a requirement that the misidentification probability must be upper bounded by δ. Furthermore, if there is a positive probability that the implemented strategy does not terminate, a straightforward application of the Markov inequality would imply $\mathbb{E}_{\boldsymbol{\mu}}[\tau] = \infty$, which is undesirable. The δ-PAC strategies (Definition 2) formally characterize these requirements, ensuring that the strategy not only stops almost surely but also identifies the best arm correctly within the specified confidence level.

Definition 2. *A δ-PAC strategy with stopping rule τ and decision rule $\hat{\imath}$ is a strategy such that for any $\boldsymbol{\mu} \in \Lambda$, $\mathbb{P}_{\boldsymbol{\mu}}[\tau < \infty] = 1$ and $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})] \leq \delta$.*

The instance-specific lower bound, derived by [29], is presented in Theorem 2. The lower bounds for general pure exploration problems are referred to in the second paper of this thesis [78]. In Theorem 2, Σ represents the $K-1$ simplex, and $\mathrm{Alt}(\boldsymbol{\mu})$ denotes the set of parameters $\boldsymbol{\lambda} \in \Lambda$ for which $1(\boldsymbol{\lambda}) \neq 1(\boldsymbol{\mu})$.

Theorem 2. *Let $\delta \in (0,1)$ and $\boldsymbol{\mu} \in \Lambda$. For any δ-PAC strategy,*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^*(\boldsymbol{\mu}) d(\delta, 1-\delta),$$

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega} \in \Sigma} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\mu_k, \lambda_k).$$

Note that $d(\delta, 1-\delta) \approx \log(1/\delta)$ as $\delta \to 0$. Hence, (2) yields that

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(\frac{1}{\delta})} \geq T^*(\boldsymbol{\mu}).$$

Proof. Consider a δ-PAC strategy. Let $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$. According to Definition 2, τ is almost surely finite, and Lemma 1 directly implies that

$$\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_k(\tau)]\, d(\mu_k, \lambda_k) \geq d(\mathbb{E}_{\boldsymbol{\mu}}[Z], \mathbb{E}_{\boldsymbol{\lambda}}[Z]), \tag{2.3}$$

where Z can be any $\mathcal{F}_\tau$-measurable random variable bounded in $[0,1]$. With the choice, $Z = \mathbb{1}\{\hat{\imath} = 1(\boldsymbol{\lambda})\}$, the definition of δ-PAC strategy and $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$ imply that the right-hand side of inequality (2.3) is $d(\mathbb{E}_{\boldsymbol{\mu}}[Z], \mathbb{E}_{\boldsymbol{\lambda}}[Z]) \geq d(\delta, 1-\delta)$. (2.3) holds for any $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$. Thus,

$$\begin{aligned} d(\delta, 1-\delta) &\leq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{k=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_k(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} d(\mu_k, \lambda_k) \\ &\leq \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sup_{\boldsymbol{\omega} \in \Sigma} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\mu_k, \lambda_k). \end{aligned} \tag{2.4}$$

This completes the proof. □

2.3 Fixed Budget Best Arm Identification

In contrast to the previous two objectives, the instance-specific lower bound in fixed budget best arm identification remains widely open. [29] conjectured that for any *consistent* strategy, defined in Definition 3, it holds that

$$\liminf_{T \to \infty} \frac{T}{\log(1/\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})])} \geq \left(\max_{\boldsymbol{\omega} \in \Sigma} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\lambda_k, \mu_k) \right)^{-1}. \tag{2.5}$$

Definition 3. *A strategy is consistent if* $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})] \to 0$ *as* $T \to \infty$.

Although [31] showed that (2.5) is achieved by a static strategy whose allocation on arm draws is the maximizer in (2.5), which is a function of $\boldsymbol{\mu}$, the parameter $\boldsymbol{\mu}$ is initially unknown. One may hope a tracking strategy can match (2.5) like fixed confidence best arm identification. However, as argued in section 1.1 in [41], estimating the maximizer requires a non-negligible cost for exploration, which is linear to T. Recently, [6,19] leveraged the worst-case lower bound in [11] to show that (2.5) cannot be reached by a single algorithm for all instances.

In the fourth paper of this thesis [76], we answer the tight lower bound in two-armed bandits and show that uniform sampling is the matching strategy. Also, the fifth paper in this thesis [79] proves an alternative lower bound for the consistent strategies, presented in (2.6).

$$\liminf_{T\to\infty} \frac{T}{\log(1/\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})])} \geq \left(\inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \max_{\boldsymbol{\omega}\in\Sigma} \sum_{k=1}^{K} \omega_k d(\lambda_k, \mu_k) \right)^{-1}. \tag{2.6}$$

Note that a simple application of the minimax inequality (see e.g. [9]) yields that

$$\inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \max_{\boldsymbol{\omega}\in\Sigma} \sum_{k=1}^{K} \omega_k d(\lambda_k, \mu_k) \geq \max_{\boldsymbol{\omega}\in\Sigma} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\lambda_k, \mu_k)$$

Since $\mathrm{Alt}(\boldsymbol{\mu})$ is not a convex domain, one cannot directly apply Sion's minimax theorem [68] to interchange max and inf. Consequently, our lower bound (2.6) cannot recover (2.5) from [29]. Furthermore, it is highly unlikely that (2.6) is tight, given that it doesn't even hold for two-armed bandits.

Chapter 3

Summaries of Papers

This chapter provides summaries of the papers, including brief introductions to their respective models and research objectives.

Paper I: Optimal Algorithms for Multiplayer Multi-Armed Bandits

Paper I is from the following draft, with [58] serves as the preliminary version.

- Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo, "Optimal Algorithms for Multiplayer Multi-Armed Bandits," In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Summary. In the context of Multiplayer Multi-Armed Bandits (MMAB), we consider M individual players or decision-makers. During each round, every player: (i) chooses an arm from the set $\mathcal{K} = \{1, \ldots, K\}$, (ii) obtains feedback regarding the chosen arm, and (iii) might communicate with adjacent players. For simplicity, we posit that in round t, if arm k is selected, the possible reward obtained is an independent random variable $X_k(t)$, following a Bernoulli distribution with mean μ_k. Additionally, we assume the average rewards $\mu = (\mu_1, \ldots, \mu_K)$ are ordered such that $\mu_1 > \mu_2 > \ldots > \mu_K$. Recently, MMAB problems have garnered significant interest. This paper delves into the two primary MMAB problems: (A) MMAB with collisions, inspired by the challenges in radio channel assignment in cognitive radios [34], and (B) MMAB without collisions, driven by sequential decision-making in social networks [44].

(A) Multiplayer MAB with collisions. In this model, a player chooses an arm and only receives its corresponding reward if no other player has made the same selection. Specifically, during round t, if a player chooses k, she can determine (1) if her choice conflicts with other players, and (2) $X_k(t)$ if there's

no conflict. This feedback scenario is known as *collision sensing* as per [8]. The players don't communicate directly; they only detect other players through collision experiences. A policy π dictates the arm each decision-maker chooses in each round. We focus on distributed policies where each decision-maker independently determines the arm to choose. This selection relies on the information available to the decision-maker: past collision and reward observations. The arm chosen by decision-maker i in round t under policy π is represented as $k_i^\pi(t)$.

Regret and its lower bound in (A). The greatest expected reward that can be gathered in each round is $\sum_{k=1}^{M} \mu_k$ (when the M top arms are played). Therefore, the regret of a policy π up to round T is defined as:

$$R^\pi(T) = T\sum_{k=1}^{M}\mu_k - \sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}[\mu_{k_i^\pi(t)}\mathbf{1}\{\{k_i^\pi(t) \neq k_j^\pi(t), \forall j \neq i\}\}].$$

As seen in classic bandit literature [43], a policy π is considered *uniformly* good if its regret satisfies $R^\pi(T) = o(T^\alpha)$ for all $\alpha > 0$ for any possible μ. From [4], we know that any uniformly good policy π, whether centralized or not, satisfies:

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq C(\mu) := \sum_{k>M} \frac{\mu_M - \mu_k}{\mathrm{kl}(\mu_k, \mu_M)}, \tag{3.1}$$

where $\mathrm{kl}(a, b)$ denotes the KL divergence between two Bernoulli distributions with respective means a and b. This result is a straightforward extension of the classic result derived by [43]. [4] also introduces a centralized policy that achieves the above asymptotic regret lower bound.

State-of-the-art algorithm in (A). In a recent study [8], the authors propose SIC-MMAB, an algorithm that uses collisions as a communication tool and satisfies the following regret:

$$\begin{aligned} R^{\mathrm{SIC}}(T) \leq & c_1 \sum_{k>M} \min\left\{\frac{\log T}{\mu_M - \mu_k}, \sqrt{T\log T}\right\} \\ & + c_2 KM\log T + c_3 KM^3 \log^2\left(\min\left\{\frac{\log T}{(\mu_M - \mu_{M+1})^2}, T\right\}\right), \end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$. While the regret of SIC-MMAB increases logarithmically with the time horizon, it doesn't match the regret lower bound (3.1). Moreover, SIC-MMAB needs to know the time horizon beforehand. It also requires complex communication phases (players must share their estimates of the arms' mean rewards), and consequently, the number of collisions used for communication increases with T.

Our contributions in (A). We introduce DPE1 (Decentralized Parsimonious Exploration), a simple policy that achieves the asymptotic fundamental regret limit (3.1). The policy is based on the observation that in a MAB problem where the decision-maker selects M arms in each round (a model known as MAB with multiple plays [4]), an optimal algorithm involves playing the $(M-1)$ best empirical arms and exploring the remaining arm according to an optimal index policy, like KL-UCB [28]. This observation that such *parsimonious* exploration is sufficient was already noted and utilized in [16] for learning -to-rank algorithms. It proves useful in the design of a decentralized MMAB algorithm: indeed, it suggests that exploration can be solely performed by a single player, the so-called *leader*. The leader maintains the set of the M best empirical arms based on the rewards she has received for the different arms. The other players, known as the *followers*, simply need to play these best empirical arms greedily. For this purpose, the leader just needs to inform the followers when the set of the M best empirical arms changes – and it can be done using collisions as proposed in [8]. Our finite-time analysis of the regret of DPE1 reveals that: for all $T \geq 3$ and any $0 < \delta < \delta_0 = \min_{1 \leq k \leq K-1} \frac{\mu_k - \mu_{k+1}}{2}$:

$$R^{\mathrm{DPE1}}(T) \leq \sum_{k>M} \frac{\mu_M - \mu_k}{\mathrm{KL}(\mu_k + \delta, \mu_M - \delta)} f(T) + K^2 M^2 \left[\frac{1}{K-M} + 284 K^{1/2} M (7 + \delta^{-2}) \right],$$

where $f(T) = \log(T) + 4 \log\log(T)$. Notably, by first letting T approach ∞ and then δ approach 0, the above result implies that DPE1 is asymptotically optimal: its regret matches the regret lower bound (3.1). DPE1 achieves the regret of the best possible centralized algorithm. Furthermore, DPE1 is simpler than SIC-MMAB, as the leader only needs to communicate the indexes of the best empirical arms, when they change. Indeed, the expected number of collisions used for communication – equivalently the number of communication bits (one might view a collision as communicating one bit) is finite (it is upper bounded by $K^2 M^2 \left[\frac{1}{(K-M)} + 242 K^{1/2} (7 + \delta^{-2}) \right]$).

(B) Multiplayer MAB without collisions. In this model, different players can choose the same arm without collisions. If a player selects arm k during round t, she receives the reward $X_k(t)$[1]. Here, the players represent the vertices of a *communication* graph $G = (V, E)$. After each round, a player can communicate with her neighbors in G. Recent studies [44, 45, 51] assume that players can communicate real numbers to their neighbors in each round. We investigate a more realistic scenario where players can only send a finite number of bits per

[1] We assume that when two players select the same arm, they receive the same random rewards for simplicity. Nonetheless, they could collect stochastically independent rewards; this would not affect the analysis of the average regret.

round. Similar to the model with collision, we are interested in distributed arm selection policies. Under such a policy π, a player i selects arm $k_i^\pi(t)$ in round t and designs messages to send to neighbors based on her past observations (the collected rewards and the messages received from her neighbors).

Regret and its lower bound in (B). The maximum expected reward that can be collected in a single round is $M\mu_1$. Therefore, the regret of a policy π up to round T is defined as:

$$R^\pi(T) = MT\mu_1 - \sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{E}[\mu_{k_i^\pi(t)}].$$

The regret of any uniformly good policy (whether centralized or not) should satisfy:

$$\liminf_{T\to\infty}\frac{R^\pi(T)}{\log(T)} \geq C_1(\mu) := \sum_{k>1}\frac{\mu_1-\mu_k}{\mathrm{KL}(\mu_k,\mu_1)}. \tag{3.2}$$

Indeed, the above asymptotic regret corresponds to the best possible regret of a single player.

State-of-art algorithm in (B). In [51], the authors introduce Distributed Delayed UCB (DD-UCB), an algorithm that merges UCB and a consensus algorithm. The latter is designed to allow all players to share similar estimates of the mean rewards of the arms, and requires each player to send a few real numbers to her neighbors in each round. DD-UCB provides the following finite-time regret guarantee[2]:

$$R^{\mathrm{DDUCB}}(T) \leq \; c_1\sum_{k>1}\frac{\log(MT)}{\mu_1-\mu_k} + c_2 M\log(M)\sum_{k>1}(\mu_1-\mu_k),$$

for some constants $c_1, c_2 > 0$. DD-UCB has the same issues as SIC-MMAB for the MMAB with collisions: its regret does not match the regret lower bound (3.2), and it requires substantial player communication.

Our contribution in (B). We propose DPE2, an algorithm based on the same *parsimonious exploration* principle as DPE1. The algorithm begins by electing a leader among the players. After this election, the leader is the only player exploring arms, again using KL-UCB indexes. The other players, the followers, simply play the best empirical arm announced by the leader. We demonstrate that the regret of DPE2 satisfies: for all $T \geq 3$ and any $0 < \delta < \delta_0$:

$$R^{\mathrm{DPE2}}(T) \leq \sum_{k>1}\frac{\mu_1-\mu_k}{\mathrm{KL}(\mu_k+\delta,\mu_1-\delta)}f(T) + 9DKM(29+K\delta^{-2}),$$

[2]This upper bound is derived for subgaussian rewards, so it is valid for Bernoulli rewards as well.

where D is the diameter of the graph G. Hence, DPE2 achieves the regret of the best possible centralized algorithm. Additionally, under DPE2, the expected number of bits used for communication is finite (it is upper bounded by $4DM^2 \log_2(M) + 8KM^2D \log_2(K)(29 + K\delta^{-2})$).

Contribution. The conceptualization and formulation of the model and problem were collaboratively developed by A. Proutiere. The proof establishment and algorithm design were carried out by the thesis author, A. Proutier, K. Ariu, and Y. Jedra. Both Y. Jedra and A. Russo conducted numerical experiments. All of the authors actively participated in the writing of the entire manuscript.

Paper II: Fast Pure Exploration via Frank-Wolfe

Paper II is from the following draft.

- Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere, "Fast Pure Exploration via Frank-Wolfe" In *The 35th Conference on Neural Information Processing Systems (NeurIPS)*., 2021.

Summary. The exploration of pure exploration in stochastic bandits [43] is a task that aims to gather information about the reward distributions of different arms with minimal arm pulls or samples. The task can encompass identifying the best arm [22, 29, 53], the top-m arms [80], all ϵ-good arms [52], or a set of arms whose expected rewards surpass a specific threshold [48]. In order to decrease the sample complexity of such a task, the learner must maximize the utilization of available information about reward distributions, which usually comes in the form of known structural properties of the set of expected rewards. Exploiting certain structures such as unimodal, Lipschitz, convex, and linear has been thoroughly researched in the regret minimization setting [15], but less so in the pure exploration framework, where the focus has largely been on linear structures [25, 33, 38, 64, 69, 72, 81].

The paper explores a generic learning problem proposed in [21] that includes the pure exploration tasks mentioned above, with or without structure. It involves K arms whose reward distributions $(\nu_1, \ldots, \nu_K)$ originate from a one-dimensional exponential family and have unknown means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. The parameter $\boldsymbol{\mu}$ is known to be a part of $\Lambda \subset \mathbb{R}^K$, the set of possible instances. For each $\boldsymbol{\mu} \in \Lambda$, it is assumed that there is a unique true answer $i^*(\boldsymbol{\mu})$ that is a part of the finite set $\mathcal{I}$ of possible answers [20]. Pure exploration tasks in the *fixed confidence* setting are considered, where the learner aims to discover $i^*(\boldsymbol{\mu})$ with a certain level of confidence $1 - \delta$, for some $\delta \in (0, 1)$. The learner's strategy is characterized by (i) an adaptive sampling rule that dictates the sequence of arm pulls, (ii) a stopping rule that defines τ, the round where the learner decides to stop pulling arms based on the data gathered so far, and (iii) a decision rule that specifies

her answer. The objective is to formulate a δ-PAC strategy that minimizes the expected sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$, and outputs the correct answer with a probability of at least $1-\delta$ for any $\boldsymbol{\mu} \in \Lambda$.

By employing the same arguments as those used in [29] for classical MAB problems, a lower bound of the expected sample complexity that any δ-PAC strategy satisfies can be derived. This lower bound is given by $T^*(\boldsymbol{\mu})\mathrm{kl}(\delta, 1-\delta)$, where the characteristic time $T^*(\boldsymbol{\mu})$ is defined through the following optimization problem:

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega} \in \Sigma} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\mu_k, \lambda_k), \tag{3.3}$$

where Σ is the $(K-1)$-dimensional simplex, $\mathrm{Alt}(\boldsymbol{\mu})$ is the set of *confusing* parameters $\boldsymbol{\lambda} \in \Lambda$ such that $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$, $\mathrm{kl}(a, b)$ is the KL divergence between two Bernoulli distributions of means a and b, and $d(\mu_k, \lambda_k)$ denotes the KL divergence of arm-k reward distributions under parameters $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$. A solution $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ of (3.3) can be interpreted as an optimal *allocation*, in the sense that pulling each arm i a proportion of round equal to $\omega_i^*(\boldsymbol{\mu})$ (in expectation) constitutes an optimal sampling rule.

Most existing algorithms that achieve asymptotically (as δ approaches 0) minimal sample complexity leverage a Track-and-Stop (TaS) framework [29]. In each round t, they plug $\hat{\boldsymbol{\mu}}(t)$, the estimated expected arm rewards, into the lower bound optimization problem (3.3), and track the allocation $\boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}(t))$. However,as previously noted in [53], the primary limitation of the Track-and-Stop framework is its requirement for recurrent access to an Oracle capable of solving (3.3). While (3.3) is a concave program, it can become challenging to solve depending on the underlying structure Λ. Particularly for complex structures, the identification of the most confusing parameters leading to the objective function $\inf_{\boldsymbol{\lambda} \in \mathrm{Alt}()} \sum_{k=1}^{K} \omega_k d(\mu_k, \lambda_k)$ can be difficult. In the following, we summarize our contribution of this paper.

(a) The paper proposes an online iterative method to approximate the optimal allocation of arm pulls, as an alternative to solving (3.3) in each round as in the TaS framework. Specifically, it introduces Frank-Wolfe-based Sampling (`FWS`), a computationally efficient algorithm that relies on a single iteration of the Frank-Wolfe (FW) algorithm applied to (3.3) instantiated at $\hat{\boldsymbol{\mu}}(t)$ in each round.

(b) For a broad class of pure exploration problems with or without structure, an upper bound of the expected sample complexity of `FWS` for any certainty level δ is derived. It is shown that this bound matches the lower bound $T^*(\boldsymbol{\mu})\mathrm{kl}(\delta, 1-\delta)$ asymptotically as δ goes to 0.

(c) The performance of `FWS` is demonstrated on various pure exploration problems, including the identification of the best arm in unstructured, linear, and Lipschitz bandits. In all tested scenarios, `FWS` matches the performance of the best existing algorithms despite its simplicity.

The use of the FW algorithm was suggested in [29] for the problem of best arm identification in unstructured bandits. In this context, FW iterations take a very simple and intuitive form (see [53]). The corresponding sampling rule, referred to as Best Challenger in [29], leads to algorithms with remarkably low sample complexity empirically, sometimes even lower than that of TaS algorithms that solve (3.3) in each round. However, as discussed in [53], the analysis of FW-type sampling rules and their convergence have remained elusive. To design the `FWS` algorithm, a simple variant of the FW algorithm is devised that yields a sampling rule whose sample complexity can be analyzed. The asymptotic optimality of this variant is confirmed, as well as its empirical superiority, not only for the case of best arm identification in unstructured bandits as predicted by [29], but also for a wide class of pure exploration problems. This analysis is believed to provide interesting solutions to the three significant obstacles needed to devise and analyze a FW-type sampling rule: (i) the objective function in (3.3) is not smooth; (ii) its curvature becomes infinite in general close to the boundary of Σ; and (iii) the estimate $\hat{\boldsymbol{\mu}}(t)$ is evolving and might be far from $\boldsymbol{\mu}$.

Contribution. The author of the thesis formulated the research question. The algorithm design and the performance guarantees are developed through the discussion between the author of the thesis and A. Proutiere. R-C. Tzeng contributed to the numerical result of this manuscript. All the authors contributed to the writing of the entire manuscript.

Paper III: Closing the Computational-Statistical Gap in Best Arm Identification for Combinatorial Semi-bandits

Paper III is from the following paper.

- Ruo-Chun Tzeng, Po-An Wang, Alexandre Proutiere, and Chi-Jen Lu, "Closing the Computational-Statistical Gap in Best Arm Identification for Combinatorial Semi-bandits" In *The 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Summary. We presents an effective method to create statistically optimal algorithms for active learning tasks. This method involves a two-step procedure. The first step involves deriving tight information-theoretical fundamental limits that a wide range of learning algorithms adhere to, through change-of-measure arguments. These limits are often expressed as the solution to an optimization problem, referred to as the *lower-bound problem*. The solution to this problem

outlines the optimal exploration process, characterizing the adaptive sampling rule that any statistically optimal algorithm should implement. The second step involves designing the learning algorithm in a way that its exploration process approaches the solution of the lower-bound problem. This method has proven to be successful for simple learning tasks such as regret minimization or best-arm identification with fixed confidence in classical stochastic bandits [27, 29, 42], but also in bandits whose arm-to-average reward function satisfies simple structural properties (e.g., Lipschitz, unimodal) [49, 78].

This method also offers an intuitive approach to examining the computational-statistical gap for active learning tasks [36]. If it is feasible to solve the lower-bound problem in polynomial time, it could pave the way for the development of learning algorithms that are both statistically optimal and computationally efficient. However, at present, the computational complexity of the lower-bound problem remains largely uncharted territory, except for straightforward learning tasks. For instance, when identifying the best policy in tabular Markov Decision Processes, the lower-bound problem is non-convex and its complexity and approximability remain ambiguous [2, 3].

In this paper, we utilize the previously mentioned two-step procedure to evaluate the computational-statistical gap for identifying the best arm in combinatorial semi-bandits under a fixed confidence setting. We demonstrate that this gap essentially does not exist, a conclusion that aligns with previous conjectures. Specifically, we introduce an algorithm that has three key features: (i) it operates within polynomial time, (ii) its sample complexity asymptotically aligns with the fundamental limits in the high confidence regime, and (iii) in the moderate confidence regime, its sample complexity is at most polynomial. Following a formal introduction to combinatorial semi-bandits, we delve into a detailed discussion of our contributions and methodologies.

Best Arm Identification in Combinatorial Semi-bandits. In the context of combinatorial semi-bandits [12, 17], a learner sequentially chooses an action from a combinatorial set, $\mathcal{X}$, which is a subset of $\{0,1\}^K$. During each round t, when the action $\boldsymbol{x}(t)$ is selected, the environment generates a K-dimensional vector $\boldsymbol{y}(t)$, assumed to follow a Gaussian distribution. The learner then receives the reward vector $\boldsymbol{x}(t) \odot \boldsymbol{y}(t)$, where $\odot$ represents the element-wise product. In simpler terms, the learner observes the individual reward $y_k(t)$ of arm k only if this arm is selected in round t, that is, $x_k(t) = 1$. The parameter $\boldsymbol{\mu}$, which characterizes the average rewards of the various arms, is initially unknown. The learner's objective is to identify the best action $\boldsymbol{i}^*(\boldsymbol{\mu})$ with a certain level of confidence $1-\delta$, while also minimizing the expected number of rounds needed. We assume that the best action is unique, and we denote Λ as the set of parameters that meet this assumption. The learner's strategy comprises three elements: (i) a sampling rule that dictates the sequence of selected actions, (ii) a stopping time τ that marks

the final round of interaction with the environment, and (iii) a decision rule that specifies the action $\hat{\boldsymbol{\imath}}$, believed to be optimal based on the data accumulated until τ.

Sample Complexity Lower-Bound Problem. Consider the set of δ-PAC algorithms such that for any $\boldsymbol{\mu} \in \Lambda$, the best action is identified correctly with probability at least $1-\delta$. Our goal is to find a δ-PAC algorithm with minimal expected sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$. To this end, we can use classical change-of-measure arguments [29] to derive a lower bound of the expected sample complexity that any δ-PAC algorithm must satisfy. This lower bound is given by $\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^*(\boldsymbol{\mu})\mathrm{KL}(\delta, 1-\delta)$. The characteristic time $T^*(\boldsymbol{\mu})$ is defined as the value of the following problem:

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega} \in \Sigma} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \langle \boldsymbol{\omega}, \frac{(\boldsymbol{\mu}-\boldsymbol{\lambda})^2}{2} \rangle, \tag{3.4}$$

where $\Sigma = \{\sum_{\boldsymbol{x} \in \mathcal{X}} \omega_{\boldsymbol{x}} \boldsymbol{x} : \boldsymbol{\omega} \in \Sigma_{|\mathcal{X}|}\}$, $\mathrm{KL}(a,b)$ is the KL-divergence between two Bernoulli distributions with respective means a and b, and $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \Lambda : \boldsymbol{\imath}^*(\boldsymbol{\lambda}) \neq \boldsymbol{\imath}^*(\boldsymbol{\mu})\}$ is the set of confusing parameters. As it turns out, $T^*(\boldsymbol{\mu})$ is at most quadratic in K, and hence the sample complexity lower bound is polynomial. The right-hand-side of (3.4) is a concave program over Σ, and a point $\boldsymbol{\omega}^*$ in its solution set corresponds to an optimal allocation of action draws: an algorithm sampling actions according to $\boldsymbol{\omega}^*$ and equipped with an appropriate stopping rule would yield a sample complexity matching the lower bound. In this paper, we provide computationally efficient algorithms to solve (3.4) and show how these can be used to devise a δ-PAC best action identification algorithm with minimal sample complexity and running in polynomial time. We only assume that we have access to a computationally efficient Oracle, referred to as the `LM` (Linear Maximization) Oracle, identifying the best action should $\boldsymbol{\mu}$ be known (but for any possible $\boldsymbol{\mu}$). This assumption, made in all previous work on combinatorial semi-bandits (see e.g. [35, 57]), is crucial as indeed, if there is no computationally efficient algorithm solving the offline problem $\arg\max_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{x}, \boldsymbol{\mu} \rangle$ with known $\boldsymbol{\mu}$, there is no hope to solve its online version with unknown $\boldsymbol{\mu}$ in a computationally efficient manner. The assumption holds for a large array of combinatorial sets of actions [63], including m-sets, matchings, (source–destination)-paths, spanning trees, matroids (refer to [18] for a thorough discussion).

The Most-Confusing-Parameter (`MCP`) algorithm. The difficulty of solving (3.4) lies in the inner optimization problem, i.e., in evaluating the objective function:

$$F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \langle \boldsymbol{\omega}, \frac{(\boldsymbol{\mu}-\boldsymbol{\lambda})^2}{2} \rangle = \min_{\boldsymbol{x} \neq \boldsymbol{\imath}^*\boldsymbol{\mu}} f_{\boldsymbol{x}}(\boldsymbol{\omega}, \boldsymbol{\mu}) \tag{3.5}$$

where $f_{\boldsymbol{x}}(\boldsymbol{\omega}, \boldsymbol{\mu}) = \inf_{\boldsymbol{\lambda} \in \mathcal{C}_{\boldsymbol{x}}} \langle \boldsymbol{\omega}, \frac{(\boldsymbol{\mu}-\boldsymbol{\lambda})^2}{2} \rangle$ and $\mathcal{C}_{\boldsymbol{x}} = \{\boldsymbol{\lambda} \in \mathbb{R}^K : \langle \boldsymbol{\lambda}, \boldsymbol{\imath}^*\boldsymbol{\mu} - \boldsymbol{x} \rangle < 0\}$. The evaluation of $F_{\boldsymbol{\mu}}(\boldsymbol{\omega})$ is necessary to solve Eq. (3.4) and to design an efficient

stopping rule. Our first contribution is the MCP (Most-Confusing-Parameter) algorithm, which can approximate $F_{\boldsymbol{\mu}}(\boldsymbol{\omega})$ for any given $\boldsymbol{\mu}$ and $\boldsymbol{\omega}$ in polynomial time. The name of algorithm name stems from the fact that by calculating $F_{\boldsymbol{\mu}}(\boldsymbol{\omega})$, we implicitly identify the most confusing parameter $\boldsymbol{\lambda}^* \in \arg\inf_{\boldsymbol{\lambda}\in \mathrm{Alt}\boldsymbol{\mu}} \langle \boldsymbol{\omega}, \frac{(\boldsymbol{\mu}-\boldsymbol{\lambda})^2}{2}\rangle$. The MCP design utilizes a Lagrangian relaxation of the optimization problem that defines $f_{\boldsymbol{x}}(\boldsymbol{\omega}, \boldsymbol{\mu})$ and leverages the fact that the Lagrange dual function linearly depends on $\boldsymbol{x}$. This linearity enables us to employ the LM Oracle. We demonstrate that computing $F_{\boldsymbol{\mu}}(\boldsymbol{\omega})$ essentially involves solving a two-player game where one player can update her strategy using the LM Oracle. We prove the following theorem, which informally quantifies the performance of the MCP algorithm:

Theorem 3. *For any $(\boldsymbol{\omega}, \boldsymbol{\mu})$, the MCP algorithm with precision ϵ and certainty parameter θ returns $\hat{F}$ and $\hat{\boldsymbol{x}}$ satisfying*

$$\mathbb{P}_{\boldsymbol{\mu}}[F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) \leq \hat{F} \leq (1+\epsilon)F_{\boldsymbol{\mu}}(\boldsymbol{\omega})] \geq 1-\theta$$

and $\hat{F} = f_{\hat{\boldsymbol{x}}}(\boldsymbol{\omega}, \boldsymbol{\mu})$. The number of calls to the LM Oracle is, almost surely, at most polynomial in K, ϵ^{-1}, and $\ln \theta^{-1}$.

The Perturbed Frank-Wolfe Sampling (P-FWS) Algorithm. The MCP algorithm enables us to solve the lower-bound problem (3.4) for any given $\boldsymbol{\mu}$. Although initially unknown, $\boldsymbol{\mu}$ could be estimated. A Track-and-Stop algorithm [29] solving (3.4) with this plug-in estimator in each round would yield asymptotically minimal sample complexity, but at a high computational cost. To overcome this issue, our P-FWS algorithm, similar to the approach in [78], performs a single iteration of the Frank-Wolfe algorithm for the program (3.4) using an estimator of $\boldsymbol{\mu}$. P-FWS employs stochastic smoothing techniques to approximate the non-differentiable objective function $F_{\boldsymbol{\mu}}$ with a smooth function. To estimate the gradient of this function, P-FWS utilizes both the LM Oracle and the MCP algorithm (specifically its second output $\hat{\boldsymbol{x}}$). Lastly, the stopping rule of P-FWS is a standard Generalized Likelihood Ratio Test (GLRT), comparing the estimated objective function to a time-dependent threshold, which also requires the MCP algorithm. We evaluate the sample and computational complexities of P-FWS. Our primary results are summarized in the following theorem.

Theorem 4. *For any $\delta \in (0,1)$, P-FWS is δ-PAC, and for any $(\epsilon, \widetilde{\epsilon}) \in (0,1)$ small enough, its sample complexity satisfies:*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \leq \frac{(1+\widetilde{\epsilon})^2}{T^*(\boldsymbol{\mu})^{-1} - \epsilon} \times H\left(\frac{1}{\delta} \cdot \frac{c(1+\widetilde{\epsilon})^2}{T^*(\boldsymbol{\mu})^{-1} - \epsilon}\right) + \Psi(\epsilon, \widetilde{\epsilon}),$$

where $H(x) = \ln(x) + \ln\ln(x)$, $c > 0$ is a universal constant, and $\Psi(\epsilon, \widetilde{\epsilon})$ is polynomial in ϵ^{-1}, $\widetilde{\epsilon}^{-1}$, K, $\|\boldsymbol{\mu}\|_\infty$, and $\Delta_{\min}^{-1}$, where $\Delta_{\min} = \min_{\boldsymbol{x} \neq \boldsymbol{i}^*\boldsymbol{\mu}} \langle \boldsymbol{i}^*\boldsymbol{\mu} - \boldsymbol{x}, \boldsymbol{\mu}\rangle$. Under P-FWS, the number of LM Oracle calls per round is at most polynomial in $\ln \delta^{-1}$ and K. The total expected number of these calls is also polynomial.

As far as we know, `P-FWS` is the first polynomial time best action identification algorithm with minimal sample complexity in the high confidence regime (when δ tends to 0). Its sample complexity is also polynomial in K in the moderate confidence regime.

Contribution. The question is formulated by the author of the thesis. One key idea is developed through the discussion between R-C. Tzeng and C-J. Lu. The other key idea and the complete proof establishment are developed by R-C. Tzeng and the author of the thesis. R-C. Tzeng, A. Proutiere, and the author of the thesis actively contributed the writing of the entire manuscript.

Paper IV: On Universally Optimal Algorithms for A/B Testing

Paper IV is from the following draft, with [76] serves as the preliminary version.

- Po-An Wang, Kaito Ariu, and Alexandre Proutiere, "On Universally Optimal Algorithms for A/B Testing," In *The 41st International Conference on Machine Learning (ICML)*, 2024

Summary. We investigate of Fixed-Budget Best-Arm Identification (FB-BAI) in Bernoulli reward-based stochastic multi-armed bandits. The learner in this scenario sequentially selects an arm and observes a randomly generated reward according to the corresponding distribution. The arms' expected rewards are initially unknown. The learner has a fixed budget of $T \in \mathbb{N}$ pulls or samples, and after collecting these samples, she must identify the arm she believes has the highest average reward. For any $k \in [K] \coloneqq \{1, \dots, K\}$, $\mu_k \in (0,1)$ represents the unknown average reward of arm k. We assume that the best arm is unique and define the parameter set of the average rewards as $\Lambda = \{\boldsymbol{\mu} \in (0,1)^K : \exists k : \mu_k > \mu_j, \forall j \neq k\}$. A strategy for FB-BAI consists of a *sampling rule* and a *decision rule*. The sampling rule determines the arm $A_t \in [K]$ to be explored in round t, based on past observations. The corresponding observed reward is $X_t \in \{0,1\}$. The arm A_t selected in round t is $\mathcal{F}_t$ measurable where $\mathcal{F}_t$ denotes the σ-algebra generated by the set of random variables $\{A_1, X_1, \dots, A_{t-1}, X_{t-1}\}$. After T rounds, the decision rule returns an answer $\hat{\imath} \in [K]$, which is $\mathcal{F}_T$ measurable. The goal is to identify a strategy that minimizes the error probability defined as

$$p_{\boldsymbol{\mu},T} \coloneqq \mathbb{P}_{\boldsymbol{\mu}} \left[\hat{\imath} \neq 1(\boldsymbol{\mu}) \right],$$

where $1(\boldsymbol{\mu}) \coloneqq \arg\max_k \mu_k$ denotes the unique best arm under the instance $\boldsymbol{\mu}$.

A naive strategy involves allocating a fixed portion of the budget to sample each arm. Once the budget is exhausted, the strategy then returns the arm with the highest empirical mean. We refer to such a strategy as a *static* algorithm (in contrast to adaptive algorithms that may select the arm to pull next based

on the rewards observed so far). An example of a static strategy is the uniform sampling strategy that distributes the budget evenly among arms. Static algorithms are well-understood and in particular, their asymptotic error rates are known [31]. Many adaptive sampling algorithms have been designed, see, e.g., [7, 26, 37, 41, 62, 79], with the hope of an improved performance compared to static algorithms. It is still unclear whether this hope can actually be fulfilled.

Despite recent research efforts, the FB-BAI problem remains largely open [60]. This contrasts with the two other classical learning tasks in stochastic bandits, namely regret minimization [43] and best arm identification with fixed confidence [29]. Indeed, for these tasks, asymptotic instance-specific performance limits and matching algorithms have been derived. In this paper, we aim at improving our understanding of the FB-BAI problem and more specifically at answering the following two natural questions, mentioned as open problems in [60].

Open problem 1. Is there an algorithm whose performance is as good as that of the uniform sampling algorithm on all instances and that strictly outperforms the latter on some instances?

Open problem 2. Can we derive an asymptotic and instance-specific error rate lower bound that (i) is satisfied by all algorithms within a wide class $\mathcal{A}$ of algorithms and that (ii) is achieved by a single algorithm in $\mathcal{A}$ on all instances? We address both open problems in the case of the FB-BAI problem with two arms (also referred to as the A/B testing problem) with Bernoulli rewards. We also provide a first set of results towards addressing these problems in the general setting with more than two arms. More precisely our contributions are as follows.

(a) For the A/B testing problem, we prove that there is no algorithm strictly outperforming the uniform sampling algorithm. To this aim, we first introduce the natural class of consistent and stable algorithms (stability here just means that the algorithm exhibits a symmetric and continuous behavior with respect to the instances). We then show that this class includes any algorithm performing as well as the uniform sampling algorithm on all instances. We finally derive an instance-specific lower bound on the error rate satisfied by any consistent and stable algorithm. As it turns out, this lower bound corresponds to the performance of the uniform sampling algorithm. Therefore, the answer to the first open problem is negative.

(b) Our analysis further provides a positive answer to the second open problem. Indeed, it yields an instance-specific error rate lower bound for the class of consistent and stable algorithms, and the performance of the uniform sampling algorithm matches this fundamental limit.

(c) For the FB-BAI problem with more than two arms, we manage to exactly

characterize the asymptotic error rate of the renowned Successive Rejects (SR) algorithm [7]. This contrasts with existing analyses of adaptive algorithms where only upper bounds of the error rate can be derived. Using this characterization, we show that, surprisingly, the uniform sampling algorithm outperforms the SR algorithm in certain instances.

Contribution. K. Ariu and the author of the thesis contributed to the problem formulations through active discussion. K. Ariu and the author of the thesis established the proof. A. Proutiere offered some ideas for the direction of the proof. The initial manuscript was primarily written by the author of the thesis and K. Ariu, with all authors actively contributing to subsequent revisions.

Paper V:

Paper V is from the following paper.

- Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere, "Best Arm Identification with Fixed Budget: A Large Deviation Perspective" In *The 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Summary. We investigate the best-arm identification issue in the fixed budget setting of stochastic bandits, commonly referred to as FB-BAI. In this scenario, a learner interacts with K distributions or arms $\nu_1, \dots, \nu_K$, defined by their unknown means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ (we limit our focus to distributions from a one-parameter exponential family). The learner sequentially pulls arms and observes samples from the corresponding distributions. Specifically, in round $t \geq 1$, she pulls an arm $A_t = k$ chosen based on previous observations and observes $X_k(t)$, a sample of a ν_k-distributed random variable. $(X_k(t), t \geq 1, k \in [K])$ are assumed to be independent across rounds and arms. After T arm draws, the learner returns $\hat{\imath}$, an estimate of the best arm $1(\boldsymbol{\mu}) := \arg\max_k \mu_k$. We assume that the best arm is unique, and denote by Λ the set of parameters $\boldsymbol{\mu}$ such that this assumption holds. The goal is to develop an adaptive sampling algorithm that minimizes the error probability $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})]$. This learning task is a key problem in stochastic bandits, and despite recent research efforts, it remains largely unresolved [60]. Particularly, researchers have yet to characterize the minimal instance-specific error probability. This contrasts with other fundamental learning tasks in stochastic bandits such as regret minimization [43] and BAI with fixed confidence [29], for which indeed, asymptotic instance-specific performance limits and matching algorithms have been derived. In FB-BAI, the error probability typically decreases exponentially with the sample budget T, i.e., it scales as $\exp(-R(\boldsymbol{\mu})T)$ where the instance-specific rate $R(\boldsymbol{\mu})$ depends on the sampling algorithm. Maximizing this rate over the set of adaptive algorithms is an open problem.

Instance-specific error probability lower bound. To estimate the maximal rate at which the error probability decays, one could apply the same strategy as that used in regret minimization or BAI in the fixed confidence setting: (i) derive instance-specific lower bound for the error probability for some notion of *uniformly good* algorithms; (ii) devise a sampling strategy mimicking the optimal proportions of arm draws identified in the lower bound. Here the notion of uniformly good algorithms is that of *consistent* algorithms. Under such an algorithm, for any $\boldsymbol{\mu} \in \Lambda$, $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} = 1(\boldsymbol{\mu})] \to 1$ as $T \to \infty$. [29] conjectures the following asymptotic lower bound satisfied by any consistent algorithm: as $T \to \infty$,

$$\frac{1}{T} \log \frac{1}{\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})]} \leq \max_{\boldsymbol{\omega} \in \Sigma} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \Psi(\boldsymbol{\lambda}, \boldsymbol{\omega}), \tag{3.6}$$

where Σ is the $(K-1)$-dimensional simplex, $\Psi(\boldsymbol{\lambda}, \boldsymbol{\omega}) = \sum_{k=1}^{K} \omega_k d(\lambda_k, \mu_k)$, $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \Lambda : 1(\boldsymbol{\mu}) \neq 1(\boldsymbol{\lambda})\}$ is the set of confusing parameters (those for which $1(\boldsymbol{\mu})$ is not the best arm), and $d(x, y)$ denotes the KL divergence between two distributions of parameters x and y. Interestingly, the solution $\boldsymbol{\omega}^* \in \Sigma$ of the optimization problem $\max_{\boldsymbol{\omega} \in \Sigma} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \Psi(\boldsymbol{\lambda}, \boldsymbol{\omega})$ provides the best static proportions of arm draws. More precisely, an algorithm selecting arms according to the allocation $\boldsymbol{\omega}^*$, i.e., selecting arm k $\omega_k^* T$ times and returning the best empirical arm after T samples, has an error rate matching the lower bound (3.6). This is a direct consequence of the fact that, under a static algorithm with allocation $\boldsymbol{\omega}$, the empirical reward process $\{\hat{\boldsymbol{\mu}}(t)\}_{t \geq 1}$ satisfies a LDP with rate function $\boldsymbol{\lambda} \mapsto \Psi(\boldsymbol{\lambda}, \boldsymbol{\omega})$, see [31] for more detail.

Adaptive sampling algorithms and their analysis. The optimal allocation $\boldsymbol{\omega}^*$ is dependent on the instance $\boldsymbol{\mu}$ and is initially unknown. We could design an adaptive sampling algorithm that (i) estimates $\boldsymbol{\omega}^*$ and (ii) follows this estimated optimal allocation. In the BAI with fixed confidence, such a tracking scheme exhibits asymptotically optimal performance [29]. However, the error made in estimating $\boldsymbol{\omega}^*$ would inevitably affect the overall error probability of the algorithm. To quantify this impact or more generally to analyze the performance of adaptive algorithms, it is crucial to understand the connection between the statistical properties of the arm selection process and the asymptotic statistics of the estimated expected rewards. Specifically, any adaptive algorithm generates a stochastic process $\{Z(t)\}_{t \geq 1} = \{(\boldsymbol{\omega}(t), \hat{\boldsymbol{\mu}}(t))\}_{t \geq 1}$. $\boldsymbol{\omega}(t) = (\omega_1(t), \ldots, \omega_K(t))$ represents the allocation realized by the algorithm up to round t ($\omega_k(t) = N_k(t)/t$ and $N_k(t)$ denotes the number of times arm k has been selected up to round t). $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \ldots, \hat{\mu}_K(t))$ denotes the empirical average rewards of the various arms up to round t. Now assuming that at the end of round T, the algorithm returns the arm with the highest empirical reward, the error probability is $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq 1(\boldsymbol{\mu})] = \mathbb{P}_{\boldsymbol{\mu}}[\hat{\boldsymbol{\mu}}(T) \in \mathrm{Alt}(\boldsymbol{\mu})]$. Assessing the error probability at least asymptotically requires understanding the asymptotic behavior of $\hat{\boldsymbol{\mu}}(t)$ as t grows large. Ideally, one would wish to establish the Large Deviation properties of the process $\{Z(t)\}_{t \geq 1}$. This task is easy for algorithms using static allocations [31],

but becomes challenging and open for adaptive algorithms. Addressing this challenge is the main objective of this paper.

In this paper, we develop and leverage tools towards the analysis of adaptive sampling algorithms for the FB-BAI problem. Our contributions are as follows.

(a) We establish a connection between the LDP satisfied by the empirical proportions of arm draws $\{\boldsymbol{\omega}(t)\}_{t\geq 1}$ and that satisfied by the empirical arm rewards. This connection holds for any adaptive algorithm. Specifically, we show that if the rate function of $\{\boldsymbol{\omega}(t)\}_{t\geq 1}$ is lower bounded by $\boldsymbol{\omega} \mapsto I(\boldsymbol{\omega})$, then that of $(\hat{\boldsymbol{\mu}}(t))_{t\geq 1}$ is also lower bounded by $\boldsymbol{\lambda} \mapsto \min_{\boldsymbol{\omega}\in\Sigma} \max\{\Psi(\boldsymbol{\lambda},\boldsymbol{\omega}), I(\boldsymbol{\omega})\}$. This result has interesting interpretations and implies the following asymptotic upper bound on the error probability of the algorithm considered: as $T \to \infty$,

$$\frac{1}{T}\log\frac{1}{\mathbb{P}_{\boldsymbol{\mu}}\left[\hat{\imath} \neq 1(\boldsymbol{\mu})\right]} \geq \inf_{\boldsymbol{\omega}\in\Sigma,\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \max\{\Psi(\boldsymbol{\lambda},\boldsymbol{\omega}), I(\boldsymbol{\omega})\}. \tag{3.7}$$

The above formula, when compared to the lower bound (3.6), quantifies the price of not knowing $\boldsymbol{\omega}^*$ initially, and relates the error probability to the asymptotic statistics of the sampling process used by the algorithm.

(b) We show that by simply applying our generic Large Deviation result, we may improve the error probability upper bounds of some existing algorithms, such as the celebrated `SR` algorithm [7]. Our result further opens up opportunities to devise and analyze new algorithms with a higher level of adaptiveness. In particular, we present `CR` (Continuous Rejects), an algorithm that, unlike `SR`, can eliminate arms in *each* round. This sequential elimination process is performed by comparing the empirical rewards of the various candidate arms using continuously updated thresholds. Leveraging the LDP tools developed in (a), we establish that `CR` enjoys better performance guarantees than `SR`. Hence `CR` becomes the algorithm with the lowest instance-specific and guaranteed error probability. We illustrate our results via numerical experiments, and compare `CR` to other BAI algorithms. This paper, therefore, makes significant strides towards understanding and improving adaptive sampling algorithms for the FB-BAI problem.

Contribution. The question is formulated by the author of the thesis and A. Proutiere. The algorithm design and the performance guarantees are developed through the discussion between the author of the thesis and A. Proutiere. R-C. Tzeng contributed to the numerical result of this manuscript. The author of the thesis and A. Proutiere contributed to the writing of the entire manuscript.

References

[1] M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

[2] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25852–25864. Curran Associates, Inc., 2021.

[3] Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes. In *Proc. of ICML*, 2021.

[4] Venkat Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, November 1987.

[5] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

[6] Kaito Ariu, Masahiro Kato, Junpei Komiyama, Kenichiro McAlinn, and Chao Qin. Policy choice and best arm identification: Asymptotic analysis of exploration sampling. *arXiv preprint arXiv:2109.08229*, 2021.

[7] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 41–53, 2010.

[8] Etienne Boursier and Vianney Perchet. SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits. In *Advances in Neural Information Processing Systems 32*, pages 12048–12057, 2019.

[9] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[10] Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[11] Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

[12] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 2012.

[13] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[14] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

[15] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Proc. of NeurIPS*, 2017.

[16] Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. Learning to rank: Regret lower bounds and efficient algorithms. *SIGMETRICS Perform. Eval. Rev.*, 43(1):231–244, June 2015.

[17] Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. In *Proc. of NeurIPS*, 2015.

[18] Thibaut Cuvelier, Richard Combes, and Eric Gourdin. Statistically efficient, polynomial-time algorithms for combinatorial semi-bandits. *Proc. of SIGMETRICS*, 2021.

[19] Rémy Degenne. On the existence of a complexity in fixed budget bandit identification. *arXiv preprint arXiv:2303.09468*, 2023.

[20] Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. In *Proc. of NeurIPS*, 2019.

[21] Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *Proc. of NeurIPS*, 2019.

[22] Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *Proc. of ICML*, 2020.

[23] Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, pages 2443–2452. PMLR, 2020.

[24] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[25] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Proc. of NeurIPS*, 2019.

[26] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

[27] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proc. of COLT*, 2011.

[28] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

[29] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[30] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

[31] Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.

[32] Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756, 2015.

[33] Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

[34] Wassim Jouini, Damien Ernst, Christophe Moy, and Jacques Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *2009 3rd International Conference on Signals, Circuits and Systems (SCS)*, pages 1–6. IEEE, 2009.

[35] Marc Jourdan, Mojmír Mutnỳ, Johannes Kirschner, and Andreas Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Proc. of ALT*, 2021.

[36] Daniel Kane, Sihan Liu, Shachar Lovett, and Gaurav Mahajan. Computational-statistical gap in reinforcement learning. In *Proc. of COLT*, 2022.

[37] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pages 1238–1246. PMLR, 2013.

[38] Zohar S Karnin. Verification based solution for structured mab problems. In *Proc. of NeurIPS*, 2016.

[39] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.

[40] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[41] Junpei Komiyama, Taira Tsuchiya, and Junya Honda. Minimax optimal algorithms for fixed-budget best arm identification. *Advances in Neural Information Processing Systems*, 35:10393–10404, 2022.

[42] Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics*, 1987.

[43] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[44] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

[45] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5239–5244. IEEE, 2018.

[46] Richard Li, Allan Jabri, Trevor Darrell, and Pulkit Agrawal. Towards practical multi-object manipulation using relational reinforcement learning. In *2020 ieee international conference on robotics and automation (icra)*, pages 4051–4058. IEEE, 2020.

[47] Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. *arXiv preprint arXiv:2011.09607*, 2020.

[48] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proc. of ICML*, 2016.

[49] Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR, 2014.

[50] Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*, pages 50–56, 2016.

[51] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4529–4540. Curran Associates, Inc., 2019.

[52] Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ϵ-good arms in stochastic bandits. In *Proc. of NeurIPS*, 2020.

[53] Pierre Ménard. Gradient ascent for active exploration in bandit problems. *arXiv*, 2019.

[54] Jianjun Ni, Yinan Chen, Yan Chen, Jinxiu Zhu, Deena Ali, and Weidong Cao. A survey on theories and applications for self-driving cars based on deep learning methods. *Applied Sciences*, 10(8):2749, 2020.

[55] Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[57] Pierre Perrault, Etienne Boursier, Michal Valko, and Vianney Perchet. Statistical efficiency of thompson sampling for combinatorial semi-bandits. In *Proc. of NeurIPS*, 2020.

[58] Alexandre Proutiere and Po-An Wang. An optimal algorithm for multiplayer multi-armed bandits. *arXiv preprint arXiv:1909.13079*, 2019.

[59] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[60] Chao Qin. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, pages 5650–5654. PMLR, 2022.

[61] Chao Qin and Wei You. Dual-directed algorithm design for efficient pure exploration. *arXiv preprint arXiv:2310.19319*, 2023.

[62] Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

[63] Alexander Schrijver et al. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer, 2003.

[64] Xuedong Shang. Linbai: Gamification of pure exploration for linear bandits. `https://github.com/xuedong/LinBAI.jl`, 2021. [Online; accessed 09-May-2021].

[65] Nícollas Silva, Heitor Werneck, Thiago Silva, Adriano CM Pereira, and Leonardo Rocha. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197:116669, 2022.

[66] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[67] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[68] Maurice Sion. On general minimax theorems. 1958.

[69] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Proc. of NeurIPS*, 2014.

[70] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[71] Koji Tabata, Junpei Komiyama, Atsuyoshi Nakamura, and Tamiki Komatsuzaki. Posterior tracking algorithm for classification bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 10994–11022. PMLR, 2023.

[72] Chao Tao, Saúl Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. In *Proc. of ICML*, 2018.

[73] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[74] Ruo-Chun Tzeng, Po-An Wang, Alexandre Proutiere, and Chi-Jen Lu. Closing the computational-statistical gap in best arm identification for combinatorial semi-bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[75] Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[76] Po-An Wang, Kaito Ariu, and Alexandre Proutiere. On uniformly optimal algorithms for best arm identification in two-armed bandits with fixed budget. *arXiv preprint arXiv:2308.12000*, 2023.

[77] Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.

[78] Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

[79] Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Best arm identification with fixed budget: A large deviation perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[80] Tengyao Wang, Nitin Viswanathan, and Sébastien Bubeck. Multiple identifications in multi-armed bandits. In *Proc. of ICML*, 2013.

[81] Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *Proc. of AISTATS*, 2018.

[82] Wei You, Chao Qin, Zihao Wang, and Shuoguang Yang. Information-directed selection for top-two algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2850–2851. PMLR, 2023.

[83] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[84] Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin. " deep reinforcement learning for search, recommendation, and online advertising: a survey" by xiangyu zhao, long xia, jiliang tang, and dawei yin with martin vesely as coordinator. *ACM sigweb newsletter*, 2019(Spring):1–15, 2019.