



Doctoral Thesis in Electrical Engineering

# Inference and Online Learning in Structured Stochastic Systems

KAITO ARIU

# Inference and Online Learning in Structured Stochastic Systems

KAITO ARIU

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 The 16th November 2023, at 2:30 p.m. in F3, Lindstedtsvägen 26, Stockholm.

Doctoral Thesis in Electrical Engineering  
KTH Royal Institute of Technology  
Stockholm, Sweden 2023

© Kaito Ariu

TRITA-EECS-AVL-2023:71  
ISBN 978-91-8040-730-4

Printed by: Universitetservice US-AB, Sweden 2023

## Abstract

This thesis contributes to the field of stochastic online learning problems, with a collection of six papers each addressing unique aspects of online learning and inference problems under specific structures. The first four papers focus on exploration and inference problems, uncovering fundamental information-theoretic limits and efficient algorithms under various structures. The last two papers focus on maximizing rewards by efficiently leveraging these structures.

The first paper addresses the complex problem of learning to cluster items based on binary user feedback for multiple questions. It establishes information-theoretical error lower bounds for both uniform and adaptive selection strategies under a fixed budget of rounds or users, and proposes an adaptive algorithm that efficiently allocates the budget. The second paper tackles the challenge of uncovering hidden communities in the Labeled Stochastic Block Model using single-shot observations of labels. It introduces a computationally efficient algorithm, Instance-Adaptive Clustering, which is the first to match instance-specific lower bounds on the expected number of misclassified items. The third paper delves into the best-arm identification or simple regret minimization problem within a Bayesian setting. It takes into consideration a prior distribution for the bandit problem and the expectation of simple regret with respect to that distribution, defining it as Bayesian simple regret. It characterizes the rate of Bayesian simple regret assuming certain continuity conditions on the prior, revealing that the leading term of Bayesian simple regret stems from parameters where the gap between optimal and suboptimal actions is less than  $\sqrt{(\log T)/T}$ . The fourth paper contributes to the fixed budget best-arm identification problem for two-arm bandits with Bernoulli rewards. It demonstrates the optimality of uniform sampling, which evenly samples the arms. It proves that no algorithm can outperform uniform sampling while being at least as good as uniform sampling for some bandit instances. The fifth paper revisits the regret minimization problem in sparse stochastic contextual linear bandits. It introduces a new algorithm, the Thresholded Lasso Bandit, which estimates the linear reward function and its sparse support, and then selects an arm based on these estimations. The algorithm achieves superior regret upper bounds compared to previous algorithms and numerically outperforms them. The sixth and final paper provides a theoretical analysis of recommendation systems in an online setting under unknown user-item preference probabilities and some structures. It derives regret lower bounds based on various structural assumptions and designs optimal algorithms that achieve these bounds. The analysis reveals the relative weights of the different components of regret, providing valuable insights into the efficient algorithms for online recommendation systems.

This thesis addresses the technical challenge of structured stochastic online learning problems, providing new insights into the power and limitations of adaptivity in these problems.

## Sammanfattning

Denna avhandling bidrar till området för stokastiska online inlärningsproblem, med en samling av sex papper som var och en behandlar unika aspekter av online inlärning och inferensproblem under specifika strukturer.

De första fyra pappren fokuserar på utforskning och inferensproblem, avslöjar grundläggande informationsteoretiska gränser och effektiva algoritmer under olika strukturer. De två sista pappren fokuserar på att maximera belöningar genom att effektivt utnyttja dessa strukturer. Det första pappret behandlar det komplexa problemet att lära sig att klustra objekt baserat på binär användarfeedback för flera frågor. Det fastställer informationsteoretiska fel nedre gränser för både uniform och adaptiv urvalsstrategier under en fast budget av rundor eller användare, och föreslår en adaptiv algoritm som effektivt allokerar budgeten. Det andra pappret tar sig an utmaningen att avslöja dolda samhällen i den märkta stokastiska blockmodellen med enstaka observationer av etiketter. Det introducerar en beräkningsmässigt effektiv algoritm, Instance-Adaptive Clustering, som är den första att matcha instansspecifika nedre gränser för det förväntade antalet felklassificerade objekt. Det tredje pappret gräver djupt i problemet med bästa armidentifiering eller enkel ångerminimering inom en Bayesiansk miljö. Det tar hänsyn till en fördelning för banditproblemet och förväntan om enkel ånger med avseende på den fördelningen, vilket definierar det som Bayesiansk enkel ånger. Det karakteriserar hastigheten för Bayesiansk enkel ånger under antagande av vissa kontinuitetsvillkor på det tidigare, vilket avslöjar att den ledande termen för Bayesiansk enkel ånger kommer från parametrar där gapet mellan optimala och suboptimala handlingar är mindre än  $\sqrt{(\log T)/T}$ . Det fjärde pappret bidrar till det fasta budget bästa arm identifieringsproblemet för två-arm banditer med Bernoulli belöningar. Det demonstrerar optimaliteten av uniform provtagning, som jämnt provtar armarna. Det bevisar att ingen algoritm kan överträffa uniform provtagning samtidigt som den är minst lika bra som uniform provtagning för vissa banditinstanser. Det femte pappret återbesöker ångerminimeringsproblemet i glesa stokastiska kontextuella linjära banditer. Det introducerar en ny algoritm, Thresholded Lasso Bandit, som uppskattar den linjära belöningsfunktionen och dess glesa stöd, och sedan väljer en arm baserat på dessa uppskattningar. Algoritmen uppnår överlägsna ånger övre gränser jämfört med tidigare algoritmer och överträffar dem numeriskt. Det sjätte och sista pappret ger en teoretisk analys av rekommendationssystem i en online miljö under okända användarobjekt preferens sannolikheter och vissa strukturer. Det härleder ånger nedre gränser baserat på olika strukturella antaganden och utformar optimala algoritmer som uppnår dessa gränser. Analysen avslöjar de relativa vikterna av de olika komponenterna i ånger, vilket ger värdefulla insikter i effektiva algoritmer för online rekommendationssystem.

Denna avhandling behandlar den tekniska utmaningen med strukturerade stokastiska onlineinlärningsproblem, och ger nya insikter i kraften och begränsningarna av anpassningsförmåga i dessa problem.

# List of Papers

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

## Paper I

Kaito Ariu, Jungseul Ok, Alexandre Proutiere, Se-Young Yun, and “Optimal Clustering from Noisy Binary Feedback,” *arXiv preprint arXiv:1910.06002*, 2019, submitted to *Machine Learning*.

## Paper II

Kaito Ariu, Alexandre Proutiere, Se-Young Yun, “Instance-Optimal Cluster Recovery in the Labeled Stochastic Block Model,” *arXiv preprint arXiv:2306.12968*, 2023, submitted to *Mathematics of Operations Research*.

## Paper III

Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin, “Rate-Optimal Bayesian Simple Regret in Best Arm Identification,” *Mathematics of Operations Research*, 2023.

## Paper IV

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.

## Paper V

Kaito Ariu, Kenshi Abe, and Alexandre Proutiere, “Thresholded Lasso Bandit,” In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

## Paper VI

Kaito Ariu, Narae Ryu, Se-Young Yun, and Alexandre Proutiere, “Regret in Online Recommendation Systems,” In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

### Conference Papers

1. Hiroaki Shiino, Kaito Ariu, Kenshi Abe, and Riku Togashi, “Exploration of Unranked Items in Safe Online Learning to Re-Rank,” In *The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (short paper)*, 2023.
2. Yuma Fujimoto, Kaito Ariu, and Kenshi Abe, “Learning in Multi-Memory Games Triggers Complex Dynamics Diverging from Nash Equilibrium,” In *The 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.
3. Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki “Last-Iterate Convergence with Full and Noisy Feedback in Two-Player Zero-Sum Games,” In *The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.
4. 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.

### Preprints

1. Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki, “Sling-shot Perturbation to Learning in Monotone Games ,” *arXiv preprint arXiv:2305.16610*, 2023, submitted.
2. Yuma Fujimoto, Kaito Ariu, and Kenshi Abe, “Memory Asymmetry Creates Heteroclinic Orbits to Nash Equilibrium in Learning in Zero-Sum Games,” *arXiv preprint arXiv:2305.13619*, 2023, submitted.
3. Masahiro Kato, Kaito Ariu, Masaaki Imaizumi, Masahiro Nomura, and Chao Qin, “Optimal Best Arm Identification in Two-Armed Bandits with a Fixed Budget under a Small Gap,” *arXiv preprint arXiv:2201.04469*, 2022.
4. 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, submitted to *Econometrica*.
5. Masahiro Kato and Kaito Ariu, “The Role of Contextual Information in Best Arm Identification,” *arXiv preprint arXiv:2106.14077*, 2021, submitted to *Journal of Machine Learning Research*.

6. Masahiro Kato, Kenshi Abe, Kaito Ariu, Shota Yasui, “A Practical Guide of Off-Policy Evaluation for Bandit Policy Evaluation,” *arXiv preprint arXiv:2010.12470*, 2020.



# Acknowledgement

Sometimes, we say that truth is stranger than fiction, but I feel like I'm living such a life myself. What surprises and delights me even more is that I've met wonderful people on this journey. I truly feel shaped by those around me.

First and foremost, my deepest gratitude goes to Alexandre Proutiere, my principal supervisor. He was always available to brainstorm problems with me, lend a hand when I was stuck, yet he respected my independence. He supported me even during the challenging times of COVID-19 through constant discussions over Zoom and chat, and when I made the decision to also work in the industry. His broad range of interests, passion for tackling challenging problems, theoretical rigor, and philosophy greatly inspired me. Without his support, I would not have accomplished even a fraction of what I have in my dissertation journey.

My co-supervisor, Mikael Johansson, always cared for me and reached out when I seemed unwell. I would like to express my deepest gratitude for the warm encouragement and support I received in my pursuit of a doctoral degree.

I would like to express my gratitude to my colleagues at KTH, especially within Alexandre's group: Yassir Jedra, Simon Lindståhl, Alessio Russo, Damianos Tranos, Filippo Vannella, and Po-An Wang. I also want to thank Rijad Alisic, Rodrigo Gonzalez, Hampei Sasahara, Takuya Iwaki, Yuchao Li, Othmane Mazhar, Jezdimir Milosevic, Xiaoqiang Ren, Ruo-Chun Tzeng, Yu Wang, Yu Xing, and Ingvar Ziemann for many interesting discussions at KTH. In particular, I would like to express my profound gratitude to Po-An and Ruo-Chun, who took the time and effort to visit Japan. Their presence not only sparked enlightening discussions, but also brought joy and vitality.

I am grateful to my collaborators: Kenshi Abe, Yuma Fujimoto, Prof. Masaaki Imaizumi, Prof. Atsushi Iwasaki, Yuu Jinnai, Prof. Junpei Komiyama, Shenyi Lu, Prof. Kenichiro McAlinn, Tetsuro Morimura, Masahiro Nomura, Naoto Ohsaka, Prof. Jungseul Ok, Chao Qin, Narae Ryu, Mitsuki Sakamoto, Hiroaki Shiino, Riku Togashi, Masatoshi Uehara, Yutaro Yamada, Shota Yasui, Takatoshi Yoshida, and Prof. Se-Young Yun. From them, I have gained deep insights into my research, specialized knowledge, and the importance of maintaining a relentless pursuit of improvement. Particularly, I owe a deep debt of gratitude to Prof. Se-Young Yun. His brilliant ideas and rigorous theoretical approach have pro-

vided me with immense stimulation, broadening my perspective on my research and giving me the strength to venture into new fields.

Special thanks go to researchers I met when I was at the University of Tokyo, who encouraged me to study in Sweden. Especially, Prof. Shinichi Nakasuka, Prof. Shinji Hara, Prof. Koji Tsumura, Prof. Ryu Funase, Prof. Takashi Tanaka, Prof. Yutaka Hori, and Prof. Takaya Inamori. My first year of study was spent at Lund University. I am thankful for the encounters there and for sending me off warmly. Much of my research was conducted at CyberAgent, where some of my collaborators are based. I am grateful for the many interactions and freedom there.

My research was partially supported by the Nakajima Foundation and the Masason Foundation.

I cannot forget to mention the invaluable moments of relaxation and joy I shared with my friends. Our outings, travels, and casual drinks were a source of immense pleasure. I look forward to creating more such memories in the future.

My family in Japan has been quietly supporting me throughout my journey, occasionally reaching out with words of encouragement and warmth. The love, passion, and sincerity that I've been able to learn from them since I was born is my greatest treasure.

Kaito Ariu,  
October 2023.

# Outline

<b>List of Papers</b>	<b>iii</b>
<b>Acknowledgement</b>	<b>vii</b>
<b>Outline</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Pure Exploration . . . . .	4
1.2 Leveraging Structure . . . . .	5
1.2.1 Leveraging Structure: Clustering . . . . .	5
1.2.2 Leveraging Structure: Sparsity . . . . .	5
1.3 Overview of the Thesis . . . . .	6
1.4 Some Technical Challenges . . . . .	7
<b>2 Summaries of Papers</b>	<b>9</b>
Paper I: Optimal Clustering from Noisy Binary Feedback . . . . .	9
Paper II: Instance-Optimal Cluster Recovery in the Labeled Stochastic Block Model . . . . .	11
Paper III: Rate-Optimal Bayesian Simple Regret in Best Arm Identification	14
Paper IV: On Uniformly Optimal Algorithms for Best Arm Identification in Two-Armed Bandits with Fixed Budget . . . . .	16
Paper V: Thresholded Lasso Bandit . . . . .	19
Paper VI: Regret in Online Recommendation System . . . . .	22
<b>References</b>	<b>29</b>
<b>Appended Papers</b>	<b>37</b>



# Chapter 1

## Introduction

Reinforcement learning [68] is a critical element in sequential decision-making within artificial intelligence and machine learning. Its applications are diverse, used in areas such as poker AI [12], sequential recommendation systems [1, 51, 54], advertisement creative selection [52], clinical treatment [9, 69], and robot arm manipulation [45]. One of the characteristics and strengths of reinforcement learning that enables these diverse applications is its adaptability. The algorithm can adjust to its environment while gathering data about it, even with limited prior knowledge. This adaptability is particularly useful in real-world applications where environmental knowledge is scarce, but data can be acquired with a relatively cheap cost [56]. However, adaptivity raises a fundamental problem of the trade-off between exploration and exploitation. Exploration aims to increase the accuracy of environmental knowledge, while exploitation aims to maximize (or minimize) the objective function, such as expected profit, based on the gathered information.

Such a trade-off is evident even in the simplest instance of the reinforcement learning problems, the *multi-armed bandit* problem. In the multi-armed bandit problem, the goal is to maximize expected rewards [49, 67]. An agent sequentially selects from a fixed set of actions, or 'arms', with each action yielding a reward sampled from a fixed, associated distribution. In each round, denoted as  $t \in [T] := \{1, 2, \dots, T\}$ , the agent selects an action  $A_t \in [K]$ . It then immediately receives a reward  $X_t$  for that action. Each action  $k \in [K]$  has an expected reward of  $\mu_k$ , and the rewards are assumed to be independently and identically distributed over the rounds  $[T]$ . The agent's selection is based on previous data  $\{A_1, X_1, A_2, X_2, \dots, A_{t-1}, X_{t-1}\}$ , indicating adaptability to the observed environment. The agent's objective is to maximize the total reward over all rounds, despite being unaware of the reward distributions, including their means. If the agent knew the values of  $(\mu_k)_{k \in [K]}$ , it could always choose  $A_t \in \arg \max_{k \in [K]} \mu_k$  for all rounds  $t \in [T]$ . This would maximize the total expected reward. However, this strategy is not feasible due to environmental uncertainty. When the mean

value of an arm is estimated from past data, it may significantly deviate from the true mean value. Such uncertainty can be reduced by increasing the number of times the arm is sampled, and confidence in that arm increases. However, focusing excessively on one arm might lead to missing out on other potentially best arms. This gives rise to a trade-off between exploration, which increases confidence in the average values for actions, and exploitation, where the agent chooses what it believes to be the best action.

For the multi-armed bandit problem, both the upper bound on the maximum reward achievable by an algorithm and the optimal algorithms to reach this bound are well-established [15, 16, 47]. Furthermore, research on structured bandits [19, 20, 23, 48, 53, 71], infinite-armed bandits [11, 76], and the incorporation of additional information, such as context, has also seen significant progress [2, 10, 18, 31].

## 1.1 Pure Exploration

In reinforcement learning, there has been a focus on problems beyond the maximization of cumulative reward. One such problem is the *pure exploration* problem [25, 26, 55], where the objective is to identify the best policy through adaptive sampling. The success of this learning approach is closely linked to a thorough understanding of the environment. In this context, the agent’s focus is solely on exploration and not on exploitation. This problem is also known as the *best-arm identification* problem [7, 13, 25] in the multi-armed bandit context. Here, the goal is to identify the action with the highest mean, denoted as  $\arg \max_{k \in [K]} \mu_k$ .

Despite the absence of trade-offs making this problem seemingly simpler, it has only recently been fully characterized in the fixed confidence setting [17, 22, 29, 37, 40, 74]. In the fixed-confidence best arm identification, the objective is to find the best arm with as small a sample size as possible while ensuring the best arm is identified with an accuracy below a given confidence value (see, e.g., [29, 49] for the detailed setting). There is also another setting for best arm identification, known as the fixed budget setting. In this setting, given a certain number of samples, the aim is to minimize the probability of incorrectly identifying the best arm. While the setting appears simple, many aspects of this problem are still not well-understood [3, 21, 40, 41, 44, 49, 62, 73]. This difficulty could stem from the fact that the accuracy of learning the most efficient exploration policies often rivals the accuracy of learning about the environment itself [39]. These two learning objectives are intertwined, making them difficult to separate. Current technical tools and knowledge for their *decoupling* are extremely limited.

Therefore, a deeper understanding of pure exploration could not only open up new application possibilities but also contribute to a better understanding of algorithm design for reinforcement learning problems in general.

## 1.2 Leveraging Structure

In reinforcement learning problems, it is essential to consider the *structure* of the environment. Especially in complex problems, understanding this structure and conducting learning based on it can lead to more efficient exploration. This structure-based learning can significantly improve efficiency, enabling the discovery and application of optimal policies faster than the algorithms that ignore the structure [27,61]. Even in cases where the number of states and actions is large or uncountably infinite, considering the structure can be a means to make learning possible.

Moreover, learning that takes into account the structure also provides a way to understand and manage the uncertainty of the environment. This could potentially maximize the utilization of information obtained in the learning process and enhance robustness. As a result, the agent might be able to respond more quickly and effectively to changes or non-stationarity in the environment.

Therefore, considering the structure of the environment can lead to the development of efficient learning algorithms in reinforcement learning.

### 1.2.1 Leveraging Structure: Clustering

The clustered structure is one such example, where data naturally segregates into distinct groups or *clusters*. Data points within the same cluster bear similarity, while those from different clusters tend to be dissimilar. *Clustering* [64,66] is a method employed to discover these cluster structures. It involves analyzing the dataset and grouping data points based on their similarity, a process that falls under unsupervised learning, as it doesn't require pre-labeled data.

In the realm of reinforcement learning, we can optimize learning by understanding the clustered structure and clustering the environment's states and the agent's actions. For instance, by categorizing similar states into one cluster, the agent can learn the optimal action for each cluster. This approach effectively reduces the effective state and action spaces' size, potentially enhancing the learning efficiency [6,35,79].

However, clustering comes with various challenges, such as determining the appropriate similarity measure and the number of clusters. Such difficulties are further compounded when trying to apply them to reinforcement learning problems.

### 1.2.2 Leveraging Structure: Sparsity

Another notable structure is the *sparsity* of features. This is particularly effective when dealing with high-dimensional feature maps in reinforcement learning problems [32]. By leveraging sparsity, the agent can disregard irrelevant features and concentrate on significant features that contribute to rewards and decision-

making. This approach can streamline the learning process and potentially enhance the agent’s performance.

Nevertheless, handling feature sparsity appropriately presents its own challenges. One such challenge involves correctly identifying and extracting features that contribute to rewards and decision-making, which necessitates feature selection and dimension reduction techniques. It may also give rise to computational issues [57, 59].

While high-dimensional linear regression methods (e.g., Lasso [14, 70]) are known in the supervised learning context, their applications to the online bandit problem have been limited until recently [10]. Thus, considering structures like feature sparsity is a critical aspect of reinforcement learning, but it also introduces new challenges.

### 1.3 Overview of the Thesis

This thesis contributes to various aspects of structured stochastic online learning problems. Comprising six papers, each contributes to online learning and inference problems within certain structures. The first four papers focus on pure exploration or inference problems, revealing fundamental (information-theoretic) limits and efficient algorithms under various structures. The final two papers, on the other hand, concentrate on maximizing rewards by effectively utilizing these structures. They derive information-theoretic limits and propose algorithms to tackle these problems. The contributions of each paper are outlined as follows.

The first paper [4] tackles the problem of clustering items based on binary user feedback from multiple questions. The paper establishes information-theoretical error lower bounds for both uniform and adaptive selection strategies under a fixed budget of rounds or users. An adaptive algorithm is also developed, which efficiently allocates the budget. The second paper [5] tackles the problem of uncovering hidden communities in the Labeled Stochastic Block Model [33] (LSBM) using single-shot observations of labels. This paper introduces a computationally efficient algorithm, Instance-Adaptive Clustering (IAC), which is the first to match instance-specific lower bounds on the expected number of misclassified items. The third paper, [43], delves into the best-arm identification or simple regret minimization problem within a Bayesian setting. It takes into consideration a prior distribution for the bandit problem and the expectation of simple regret with respect to that distribution, defining it as *Bayesian* simple regret. Given certain continuity conditions on the prior, it characterizes the rate of Bayesian simple regret. The paper reveals that the primary terms of Bayesian simple regret stem from parameters where the gap between optimal and suboptimal actions is less than  $\sqrt{(\log T)/T}$ . This contrasts with Bayesian regret minimization as discussed in [46]. The fourth paper, [73], makes a contribution to the problem of fixed budget best-arm identification for two-arm bandits with Bernoulli rewards. It demonstrates the optimality of *uniform sampling*, which evenly samples the

arms. Specifically, it proves that no algorithm can outperform uniform sampling while being at least as good as it for certain bandit instances. The proof's key ingredient lies in demonstrating that natural algorithms are consistent and stable. This paper provides a partial solution to the open problems presented in [62]. The fifth paper, [2], revisits the regret minimization problem in sparse stochastic contextual linear bandits. It introduces a new algorithm, the Thresholded Lasso Bandit, which estimates the linear reward function and its sparse support and then selects an arm based on these estimations. The algorithm achieves superior regret upper bounds compared to previous algorithms and numerically outperforms them. The sixth paper, [6], provides a theoretical analysis of recommendation systems in an online setting under unknown user-item preference probabilities and some structures. With  $m$  users and  $n$  items, and the constraint that an item cannot be recommended to the same user twice, the paper derives regret lower bounds based on various structural assumptions and designs optimal algorithms that achieve these bounds. The analysis reveals the relative weights of the different components of regret: (i) the component from the constraint that an item cannot be recommended to the same user twice, (ii) learning the statistical parameters, and (iii) learning the structure.

## 1.4 Some Technical Challenges

The problems considered in this thesis involve several technical challenges, including the derivation of lower bounds and theoretical analyses necessary for ensuring optimality.

**Lower bounds.** The accompanying papers derived lower bounds or fundamental limits for the error probability [4, 5, 73], Bayesian simple regret [43], and (cumulative) regret [6]. Furthermore, efficient and optimal algorithms for each problem were derived [2, 4–6, 43]. Regarding the lower bounds, all these derivations rely on the change-of-measure argument, as in [47], which are commonly used to prove lower bounds in adaptive stochastic optimization problems. This thesis addressed the technical challenge of applying the change-of-measure argument under these settings. These were challenging because separating errors in statistical parameter estimation from the error probability is not straightforward. Ideally, with known statistical parameters, lower bounds can be inferred from the classical Large Deviation Principle [24, 30]. However, the results are not easily generalizable when the algorithm's sampling rules are adaptive. The power of adaptivity posed additional technical challenges. This thesis partially derived the lower bounds on the error probability and Bayesian simple regret within a fixed-budget setting [4–6, 73], for both adaptive and non-adaptive sampling rules.

In [4] and [5], the successful derivations of the lower bounds were achieved by leveraging the fact that under a clustered structure, some information can be transferred from other items. In this case, the problem can be treated as if the statistical parameters were known, and the difficulty could be avoided by apply-

ing a change-of-measure argument that focused only on the cluster assignment error. In [73], the proof was made possible by identifying that being *stable* is necessary for any uniformly good algorithm. While the idea of imposing stability on the algorithm to achieve decoupling was in several studies [36, 78], this paper succeeded in achieving such decoupling for the fixed-budget best arm identification problem. The results are extreme in the opposite direction to the clustering case, suggesting that the error in learning the statistical parameter may be even larger than the probability of misidentification of the best arm.

The derivation of the regret lower bounds under the existence of non-repetition constraints in [6] presented difficulties, as the change-of-measure argument has to be applied to the problem with some non-asymptotic nature. In [6], the behavior of the algorithm in the two-armed bandit problem was analyzed, and the proof was made possible by defining the relationship between the two-armed bandit algorithm and the original algorithm as a simulation.

**Algorithms and its analysis.** Contributions were also in the algorithm designs and its analysis. Regarding the optimal algorithms, guarantees for the clustering algorithms in [4–6] were provided in expectation. Most clustering algorithms with instance-specific analysis only offered high probability bounds, lacking a bound in expectation [77]. This necessitated a redesign of the algorithms and the establishment of precise bounds for each step of the proof [5]. Consequently, this led to tighter regret analyses under clustered structures, as shown in [6]. In [5], an algorithm with a very small computational complexity was used, but this created correlations and made the analysis harder. Lastly, [2] succeeded in the tighter analyses by refining the idea of thresholded Lasso [80] and the regret bound of [10, 60].

## Chapter 2

# Summaries of Papers

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

### Paper I: Optimal Clustering from Noisy Binary Feedback

Paper I is from the following draft.

- Kaito Ariu, Jungseul Ok, Alexandre Proutiere, Se-Young Yun, and “Optimal Clustering from Noisy Binary Feedback,” *arXiv preprint arXiv:1910.06002*, 2019, submitted to *Machine Learning*.

**Summary.** Modern machine learning models heavily rely on large volumes of labeled data. This process can be tedious and costly, as humans are the primary source of labeling. Crowdsourcing platforms like Amazon Mechanical Turk and reCAPTCHA simplify this task. They often convert complex labeling tasks into binary questions, using user responses to label images. For instance, one can classify bird images through a series of binary questions about the bird’s appearance. However, the difficulty of images or questions can vary, leading to a higher error rate in labeling. Therefore, to build a reliable labeling system, it’s crucial to adaptively process human responses based on the difficulty of tasks. This paper studies algorithms leading to designing such systems efficiently.

The following model with a set  $\mathcal{I}$  of  $n$  items is introduced.  $\mathcal{I}$  can be partitioned into  $K$  disjoint clusters  $\mathcal{I}_1, \dots, \mathcal{I}_K$ . For each item  $i \in \mathcal{I}$ , we denote the cluster index as  $\sigma(i)$ . We assume users arrive at the system sequentially. Upon arrival, the algorithm presents a list of  $w$  items with a yes or no question selected from a set of  $L$  questions. We consider two types of algorithms: (i) one that selects the item list and the question uniformly, and (ii) one that selects them adaptively. The algorithm is given  $T \in \mathbb{N}$  as the total number of user arrivals. The parameters  $\mathbf{p} = (p_{k\ell})_{k \in [K], \ell \in [L]} \in [0, 1]^{K \times L}$  and  $\mathbf{h} = (h_i)_{i \in \mathcal{I}} \in [0.5, 1]^n$  are used to model the statistical property of the user feedback. When the  $t$ -th user arrives, she is

asked a question  $\ell_t \in [L]$  with a set of  $w$  items  $\mathcal{W}_t$ . The user then provides noisy answers for each  $i \in \mathcal{W}_t$ . If  $i \in \mathcal{I}_k$ , the answer is  $X_{i\ell_t} = +1$  with probability  $q_{i\ell_t} = h_i p_{k\ell_t} + \bar{h}_i \bar{p}_{k\ell_t}$ , and  $-1$  with probability  $1 - q_{i\ell_t}$ . The parameter  $h_i$  can be interpreted as the *hardness* of classifying the item  $i$ . If the value of  $h_i$  is closer to 0.5,  $q_{i\ell_t}$  is close to 0.5, regardless of the item cluster of  $i$  and question  $\ell_t$ . Conversely, if  $h_i$  is close to 1, the value of  $q_{i\ell_t}$  will vary significantly depending on the question  $\ell_t$ , making clustering of  $i$  easier. Under this model, the study's objective is to cluster the items with the smallest number of misclustered items.

Let for each  $k \in [K]$ ,  $\mathbf{r}_k = (r_{k\ell})_{\ell \in [L]}$  with  $r_{k\ell} = 2p_{k\ell} - 1$ .  $\|\cdot\|$  denotes the  $\ell_\infty$ -norm, i.e.,  $\|\mathbf{x}\| = \max_i |x_i|$ . The following assumptions on the model  $\mathcal{M} = (\mathbf{p}, \mathbf{h})$  are made.

$$(A1) \quad h_* = \min_{i \in \mathcal{I}} (2h_i - 1) \in (0, 1), \quad (A2) \quad \exists \eta > 0, \eta \leq p_{k\ell} \leq 1 - \eta. \\ \rho_* = \min_{k \neq k'} \min_{0 \leq c \leq \frac{1}{h_*}} \|\mathbf{c}\mathbf{r}_{k'} - \mathbf{r}_k\| > 0.$$

With (A1), we exclude the items for which clustering is not feasible. On the other hand, (A2) imposes uniformity on the statistical parameters.

A clustering algorithm is denoted as  $\pi$ . We define the clustering error rate of an item  $i \in \mathcal{I}$  as  $\varepsilon_i^\pi(n, T)$ , which is the probability that  $i$  is misclassified under the algorithm  $\pi$ , number of items  $n$ , and total number of users  $T$ . Here,  $\mathcal{E}^\pi$  signifies the set of items that are incorrectly classified by  $\pi$ . This set is constructed as  $\mathcal{E}^\pi = \cup_{k \in [K]} \mathcal{I}_k \setminus \mathcal{S}_{\gamma(k)}^\pi$ , where the output of  $\pi$  is given by  $(\mathcal{S}_1^\pi, \dots, \mathcal{S}_K^\pi)$ , and  $\gamma$  is a permutation of  $[K]$  that minimizes the number of items in  $\cup_{k \in [K]} \mathcal{I}_k \setminus \mathcal{S}_{\gamma(k)}^\pi$ . We say that an algorithm  $\pi$  is *uniformly good* if, for all  $\mathcal{M} \in \Omega$  and  $i \in \mathcal{I}$ , the error rate  $\varepsilon_i^\pi(n, T)$  approaches zero as  $T$  tends to infinity, given that  $T = \omega(n)$ . Let  $\text{KL}(a, b)$  be the Kullback-Leibler divergence between Bernoulli distributions with parameters  $a$  and  $b$ . We establish the following lower bound on the error rate of the uniform sampling algorithm.

**Theorem 1.** *For any uniform sampling algorithm, for any  $\mathcal{M} \in \Omega$  that fulfills (A1) and (A2), as  $T \rightarrow \infty$  under  $T = \omega(n)$ , for any item  $i$ , the following holds:*

$$\varepsilon_i^\pi(n, T) \geq \exp\left(-\frac{Tw}{n} \mathcal{D}_{\mathcal{M}}^U(i)(1 + o(1))\right), \\ \text{where } \mathcal{D}_{\mathcal{M}}^U(i) = \min_{k' \neq \sigma(i)} \min_{h' \in [(h_*+1)/2, 1]} \frac{1}{L} \sum_{\ell} \text{KL}(h' p_{k'\ell} + \bar{h}' \bar{p}_{k'\ell}, q_{i\ell}) > 0.$$

For the adaptive algorithm, it is not essential to provide a lower bound on  $\varepsilon_i^\pi(n, T)$ , because the algorithm would be able to adapt to the hardness of each item and optimize the overall error rate. Define such an overall error rate by  $\varepsilon^\pi(n, T) = \frac{1}{n} \sum_{i \in \mathcal{I}} \varepsilon_i^\pi(n, T)$ . For any adaptive, uniformly good algorithm, when  $T$  grows large, we obtain:

$$\varepsilon^\pi(n, T) \geq \exp\left(-\frac{Tw}{n} \tilde{\mathcal{D}}_{\mathcal{M}}^A(1 + o(1))\right),$$

$$\begin{aligned}
\text{where } \tilde{\mathcal{D}}_{\mathcal{M}}^A &= \max_{\mathbf{y} \in \mathcal{Y}(n)} -\frac{n}{Tw} \log \left( \frac{1}{n} \sum_{i=1}^n \exp \left( -\frac{Tw}{n} \mathcal{D}_{\mathcal{M}}^A(i, \mathbf{y}) \right) \right), \\
\mathcal{D}_{\mathcal{M}}^A(i, \mathbf{y}) &= \min_{j: \sigma(j) \neq \sigma(i)} \sum_{\ell} (y_{j\ell} \text{KL}(q_{j\ell}, q_{i\ell}) + y_{i\ell} \text{KL}(q_{i\ell}, q_{j\ell})), \\
\text{and } \mathcal{Y}(n) &= \left\{ \mathbf{y} \in [0, 1]^{n \times L} : \sum_{i \in \mathcal{I}, \ell \in [L]} y_{i\ell} = n \right\}.
\end{aligned} \tag{2.1}$$

In the above, the vector  $\mathbf{y}$  represents the expected frequency of each question being asked for each item. Therefore, maximizing over  $\mathbf{y}$  in equation (2.1) corresponds to the best allocation that minimizes the overall error rate.

The paper presents a uniform sampling algorithm using a K-means algorithm with the following performance guarantee.

**Theorem 2.** *Suppose  $T = \omega(n)$  and  $T = o(n^2)$ , under the algorithm,*

$$\varepsilon_i^\pi(n, T) \leq \exp \left( -C(2h_i - 1)^2 \rho_*^2 \frac{Tw}{Ln} (1 + o(1)) \right),$$

where  $C > 0$  is some universal constant and  $\rho_*$  is a constant defined in (A1).

We further design an adaptive sampling algorithm that is inspired by the lower bound. The algorithm consistently updates the estimates of model parameters and clusters. From these estimates, the algorithm further calculates the lower bounds on the probabilities of incorrectly classifying each item. The items it chooses are those with the highest lower bounds, indicating they are most likely to be misclassified. Additionally, it chooses the question that would be the most informative about these items. The numerical experiments indicate that the adaptive algorithm significantly outperforms those with a uniform strategy for selecting the tuple (item list, question), particularly when the items have heterogeneous hardness parameters.

**Contribution.** The conceptualization and formulation of the model and problem were collaboratively developed by the thesis author, A. Proutiere, and S. Yun. The proof establishment was carried out by the thesis author, J. Ok, and S. Yun. Both the thesis author and J. Ok conducted numerical experiments using synthetic data. The thesis author conducted experiments with non-synthetic data. All of the authors actively participated in the writing of the entire manuscript.

## Paper II: Instance-Optimal Cluster Recovery in the Labeled Stochastic Block Model

Paper II is from the following draft.

- Kaito Ariu, Alexandre Proutiere, Se-Young Yun, “Instance-Optimal Cluster Recovery in the Labeled Stochastic Block Model,” *arXiv preprint arXiv:2306.12968*, 2023, submitted to *Mathematics of Operations Research*.

**Summary.** Community detection or clustering is the process of grouping similar items from data, often represented by interactions between items, a concept extensively studied through the Stochastic Block Model (SBM) [34, 58]. However, the SBM may oversimplify real-world interactions, which can vary in nature, such as ratings in recommender systems or proximity levels in social networks. The Labeled Stochastic Block Model (LSBM) [33, 50, 77] addresses this by assigning labels from an arbitrary collection to describe more complex interactions, like negative edges or user ratings. This paper aims to develop a clustering algorithm employing these labels to reconstruct item clusters, with the goal of minimizing the expected number of misclassified items.

The LSBM is a model where the set  $\mathcal{I}$  of  $n$  items is randomly divided into  $K$  distinct clusters  $\mathcal{I}_1, \dots, \mathcal{I}_K$ . The cluster of each item  $i$  is denoted as  $\sigma(i)$ , and the vector  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$  represents the probabilities of items being in each cluster. In this model,  $\alpha_1, \dots, \alpha_K$  are positive constants and both  $K$  and  $\alpha$  remain unchanged as  $n$  increases. It is assumed, without loss of generality, that  $\alpha_1 \leq \dots \leq \alpha_K$ . For each edge  $(v, w)$  in  $\mathcal{I}_i \times \mathcal{I}_j$ , the learner identifies the label  $\ell$  with probability  $p(i, j, \ell)$ , independently of the labels identified on other edges. We write  $p = (p(i, j, \ell))_{i, j \in [K], \ell \in [L]}$ . The algorithm is provided with one-shot label observations (one label for each edge) and is tasked with recovering the clusters from these observations. The objective of the study is to recover these clusters with minimal clustering errors. This task becomes particularly challenging in sparse scenarios, a condition that our study specifically considers. In these scenarios,  $\bar{p} = \mathcal{O}((\log n)/n)$  and  $\bar{p}n \rightarrow \infty$  as  $n \rightarrow \infty$ , where  $\bar{p} = \max_{i, j, \ell \geq 1} p(i, j, \ell)$  is the maximum probability of observing a label different from 0.

The following assumptions are made. For all  $i, j, k \in [K]$ :

$$(A1) \quad \forall \ell \in \mathcal{L}, \quad \frac{p(i, j, \ell)}{p(i, k, \ell)} \leq \eta \quad \text{and} \quad (A2) \quad \frac{\sum_{k=1}^K (\sum_{\ell=1}^L (p(i, k, \ell) - p(j, k, \ell)))^2}{\bar{p}^2} \geq \varepsilon,$$

where  $\eta, \varepsilon > 0$  are some constants.

We denote  $\mathcal{P}^{K \times (L+1)}$  as the set of all  $K \times (L+1)$  matrices, where each row represents a probability distribution. We define the divergence  $D(\alpha, p)$  of the parameter  $(\alpha, p)$  as follows:

$$D(\alpha, p) = \min_{i, j \in [K]: i \neq j} D_{L+}(\alpha, p(i), p(j)) \quad (2.2)$$

with  $D_{L+}(\alpha, p(i), p(j))$

$$= \min_{y \in \mathcal{P}^{K \times (L+1)}} \max \left\{ \sum_{k=1}^K \alpha_k \text{KL}(y(k), p(i, k)), \sum_{k=1}^K \alpha_k \text{KL}(y(k), p(j, k)) \right\},$$

where KL is the Kullback-Leibler divergence between two label distributions, i.e.,  $\text{KL}(y(k), p(i, k)) = \sum_{\ell=0}^L y(k, \ell) \log \frac{y(k, \ell)}{p(i, k, \ell)}$ . In the above definition, the term  $D_{L+}(\alpha, p(i), p(j))$  can be interpreted as the hardness of distinguishing whether an item belongs to cluster  $i$  or cluster  $j$ . Now, consider a clustering algorithm  $\pi$ . Let  $\varepsilon^\pi(n)$  denote the number of incorrectly classified items for a given clustering algorithm  $\pi$ , and  $\mathbb{E}[\varepsilon^\pi(n)]$  is its expected value. This quantity is defined up to a permutation. Specifically, if  $\pi$  returns  $(\hat{\mathcal{I}}_k)_k$ , then  $\varepsilon^\pi(n)$  is computed as the minimum over all permutations  $\theta$  of  $|\cup_k \hat{\mathcal{I}}_k \setminus \mathcal{I}_{\theta(k)}|$ . We show the following theorem that provides a lower bound on  $\mathbb{E}[\varepsilon^\pi(n)]$ .

**Theorem 3.** *Under the assumptions (A1), (A2),  $s = o(n)$ , and  $\bar{p}n = \omega(1)$ , for any clustering algorithm  $\pi$  that satisfies the condition  $\limsup_{n \rightarrow \infty} \frac{\mathbb{E}[\varepsilon^\pi(n)]}{s} \leq 1$ , it follows that:*

$$\liminf_{n \rightarrow \infty} \frac{nD(\alpha, p)}{\log(n/s)} \geq 1. \quad (2.3)$$

The proof, adapted from [77], is based on the change-of-measure argument from online stochastic optimization [47]. The key contribution of this paper lies in proposing an algorithm, referred to as Instance-Adaptive Clustering (IAC), that achieves the lower bound. We make the following additional assumption.

$$(A3) \quad np(j, i, \ell) \geq (n\bar{p})^\kappa \text{ for all } i, j \text{ and } \ell \geq 1, \text{ for some constant } \kappa > 0.$$

The performance guarantees of IAC are given as follows.

**Theorem 4.** *Under the assumptions (A1), (A2), and (A3), if  $(\alpha, p)$  satisfies (2.3), IAC misclassifies at most  $s$  items with high probability and in expectation, i.e.,*

$$\lim_{n \rightarrow \infty} \mathbb{P}[\varepsilon^{\text{IAC}}(n) \leq s] = 1$$

and

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E}[\varepsilon^{\text{IAC}}(n)]}{s} \leq 1.$$

IAC is a two-step algorithm that initially applies a spectral clustering algorithm to provide rough estimates of the clusters and subsequently improves the cluster assignment based on the likelihood. Notably, IAC is the first algorithm to match the lower bound (2.3) in expectation, unlike the existing instance-optimal algorithm, which only provides a high probability guarantee [77]. Furthermore, IAC is computationally efficient with a complexity of  $\mathcal{O}(npolylog(n))$ . The proof of this was made possible by providing detailed probability guarantees at each step and developing analysis techniques even under the correlation in the likelihood-based cluster improvement. As suggested in [28], the difficulties associated with such proof could be circumvented by performing the initial clustering  $n$  times,

instead of just once. However, this approach would not achieve the computational complexity of  $\mathcal{O}(npolylog(n))$ .

**Contribution.** The research question was formulated through discussions between the thesis author and A. Proutiere. The author of the thesis contributed to the algorithm design and performance guarantees. A. Proutiere and S. Yun provided some key ideas for the proof. The author of the thesis conducted numerical experiments. The author of the thesis wrote the majority of the manuscript with A. Proutiere contributing to certain parts.

### Paper III: Rate-Optimal Bayesian Simple Regret in Best Arm Identification

Paper III is from the following paper.

- Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin, “Rate-Optimal Bayesian Simple Regret in Best Arm Identification,” *Mathematics of Operations Research*, 2023.

**Summary.** This paper considers the problem of finding the optimal action from a set of  $K$  actions with a fixed sample size  $T$ . In this setting, each action  $i$ , taken from the set  $[K] = \{1, 2, \dots, K\}$ , is associated with an unknown parameter  $\mu_i \in [0, 1]$ . We represent the collection of these parameters as  $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_K) \in [0, 1]^K$ . For each round  $t = 1, \dots, T$ , an adaptive algorithm selects an action  $I(t)$  from the set  $[K]$ . The algorithm then immediately observes the reward  $X_{I(t)}(t)$ , which follows a Bernoulli distribution with parameter  $\mu_{I(t)}$ . After the reward observation at round  $t = T$ , the algorithm recommends an action  $J(T)$ . We define  $i^* = \arg \max_i \mu_i$  (assuming the uniqueness) and  $\mu^* = \mu_{i^*}$  as the optimal action and its associated mean, respectively. We denote  $\mu_{\setminus i}^* = \max_{j \neq i} \mu_j$ . In this setting, we often consider the *simple regret* defined as follows:

$$R_{\boldsymbol{\mu}}(T) = \mu^* - \mathbb{E}_{\boldsymbol{\mu}}[\mu_{J(T)}], \quad (2.4)$$

which is the expected difference of the means between the optimal action and the recommended action. In the instance-specific setting, the probability of error becomes almost equivalent to simple regret as  $T$  grows large. This is because the probability of error decreases exponentially fast. The problem of minimizing (2.4) is referred to as simple regret minimization or best arm identification and is extensively studied, e.g., [7, 13, 38]. The definition of simple regret (2.4) is frequentist in the sense that the value of  $\boldsymbol{\mu}$  is fixed as a non-random vector. In this regard, we extend this concept by considering  $\boldsymbol{\mu}$  as a random vector. By computing the expectation of the frequentist regret, we introduce the term *Bayesian simple regret*, which is defined as:

$$R_H(T) = \mathbb{E}_{\boldsymbol{\mu} \sim H}[R_{\boldsymbol{\mu}}(T)], \quad (2.5)$$

where the expectation  $\mathbb{E}_{\boldsymbol{\mu} \sim H}$  is taken over the prior distribution  $H$  on  $[0, 1]^K$ . The objective of this paper is to minimize the Bayesian simple regret (2.5).

We denote the parameters set excluding  $\mu_i$  from  $\boldsymbol{\mu}$  as  $\boldsymbol{\mu}_{\setminus i}$ , which contains  $K-1$  parameters. Similarly, when considering two indices  $i, j \in [K]$ , we denote parameter set excluding  $\mu_i$  and  $\mu_j$  from  $\boldsymbol{\mu}$  as  $\boldsymbol{\mu}_{\setminus ij}$ , which contains  $K-2$  parameters. We define  $H_{\setminus i}(\boldsymbol{\mu}_{\setminus i})$  as the joint cumulative density function of the parameters in  $\boldsymbol{\mu}_{\setminus i}$ . In addition, we define  $H_i(\mu_i | \boldsymbol{\mu}_{\setminus i})$  as the conditional cumulative density function of  $\mu_i$ , given the parameter  $\boldsymbol{\mu}_{\setminus i}$ . Similarly, we define  $H_{\setminus ij}(\boldsymbol{\mu}_{\setminus ij})$  and  $H_{ij}(\mu_i, \mu_j | \boldsymbol{\mu}_{\setminus ij})$  as the joint and conditional cumulative density functions. Let  $\setminus i^*$  be the second-best arm. We made the following assumption that the derivatives of  $H_i(\mu_i | \boldsymbol{\mu}_{\setminus i})$  and  $H_{ij}(\mu_i, \mu_j | \boldsymbol{\mu}_{\setminus ij})$  exist and they are uniformly continuous:

**Assumption 1.** *Conditional probability density functions  $h_i(\mu_i | \boldsymbol{\mu}_{\setminus i})$  and  $h_{ij}(\mu_i, \mu_j | \boldsymbol{\mu}_{\setminus ij})$  exist, and they are uniformly continuous: for any  $\epsilon > 0$  there exists  $\delta = \delta(\epsilon) > 0$  such that*

$$\begin{aligned} \forall |\mu_i - \lambda_i| \leq \delta, & \quad \Rightarrow |h_i(\mu_i | \boldsymbol{\mu}_{\setminus i}) - h_i(\lambda_i | \boldsymbol{\mu}_{\setminus i})| \leq \epsilon, \\ \forall |\mu_i - \lambda_i|, |\mu_j - \lambda_j| \leq \delta, & \quad \Rightarrow |h_{ij}(\mu_i, \mu_j | \boldsymbol{\mu}_{\setminus ij}) - h_{ij}(\lambda_i, \lambda_j | \boldsymbol{\mu}_{\setminus ij})| \leq \epsilon. \end{aligned}$$

We propose a novel algorithm called the Two-Stage Exploration (TSE). This algorithm is composed of two distinct phases: the initial phase involves uniform exploration, while the subsequent phase concentrates on the most promising arm candidates identified during the first phase. We obtain a Bayesian simple regret bound for TSE.

**Theorem 5.** *Define  $T' = \frac{2qT}{K} + (1-q)T$ . Under Assumption 1, for any  $q > 0$ , the Bayesian simple regret of TSE satisfies:*

$$R_H(T) \leq \frac{1}{T'} \sum_{i=1}^K \int_{[0,1]^{K-1}} \mu_{\setminus i}^* (1 - \mu_{\setminus i}^*) h_i(\mu_{\setminus i}^* | \boldsymbol{\mu}_{\setminus i}) dH_{\setminus i}(\boldsymbol{\mu}_{\setminus i}) + o\left(\frac{1}{T}\right). \quad (2.6)$$

This theorem indicates that as  $q \rightarrow 0$ , the Bayesian simple regret bound for TSE scales accordingly:

$$R_H(T) \leq \frac{1}{T} \sum_{i=1}^K \int_{[0,1]^{K-1}} \mu_{\setminus i}^* (1 - \mu_{\setminus i}^*) h_i(\mu_{\setminus i}^* | \boldsymbol{\mu}_{\setminus i}) dH_{\setminus i}(\boldsymbol{\mu}_{\setminus i}) + o\left(\frac{1}{T}\right).$$

Furthermore, we prove the following lower bound hold for any algorithm.

**Theorem 6.** *Under Assumption 1, for any algorithm,*

$$R_H(T) \geq \frac{1}{4.8T} \sum_{i=1}^K \int_{[0,1]^{K-1}} \mu_{\setminus i}^* (1 - \mu_{\setminus i}^*) h_i(\mu_{\setminus i}^* | \boldsymbol{\mu}_{\setminus i}) dH_{\setminus i}(\boldsymbol{\mu}_{\setminus i}) - o\left(\frac{1}{T}\right).$$

Thus, we characterize the optimal rates of the Bayesian simple regret, up to a factor of 4.8.

We discovered a key insight from the proof. The leading term of the Bayesian simple regret originates from the region with a smaller gap:  $\mu_i - \mu_{\setminus i}^* = o(\sqrt{(\log T)/T})$ . This contrasts with the result by [46], who proves that for an asymptotically optimal algorithm, the Bayesian (cumulative) regret scales as

$$\frac{(\log T)^2}{2} \sum_{i=1}^K \int_{[0,1]^{K-1}} h_i(\mu_{\setminus i}^* | \boldsymbol{\mu}_{\setminus i}) dH_{\setminus i}(\boldsymbol{\mu}_{\setminus i}) + o((\log T)^2).$$

**Contribution.** All authors contributed to the formulation of the problem through discussions. J. Komiyama was the main driver for the proof of the upper/lower bounds. The author of the thesis contributed to some part of the proof. J. Komiyama wrote a large part of the initial manuscript. All of the authors actively participated in the revision of the entire manuscript.

## Paper IV: On Uniformly Optimal Algorithms for Best Arm Identification in Two-Armed Bandits with Fixed Budget

Paper IV is from the following draft.

- 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.

**Summary.** In this study, we examined the problem of identifying the best arm with a fixed budget in two-arm bandits with stochastic Bernoulli rewards. An agent pulls an arm sequentially to observe a randomly generated reward according to the associated distribution. Initially, the expected rewards of the arms are unknown. The agent has a predetermined budget of  $T \in \mathbb{N}$  samples. After collecting these samples, the agent must recommend which arm they believe has the highest average reward. For any  $k \in \{1, 2\}$ ,  $\mu_k \in (0, 1)$  represents the unknown average reward of arm  $k$ . We define the parameter set of the mean rewards as  $\Lambda = \{\boldsymbol{\mu} \in (0, 1)^2 : \mu_1 \neq \mu_2\}$ . A strategy for identifying the best arm with a fixed budget consists of a sampling rule and a decision rule. The sampling rule determines which arm  $A_t \in \{1, 2\}$  should be explored in round  $t$ , based on previously observed rewards. The observed reward corresponding to this is  $X_t \in \{0, 1\}$ .  $\mathcal{F}_t$  symbolizes the  $\sigma$ -algebra created by the set of random variables  $\{A_1, X_1, \dots, A_{t-1}, X_{t-1}\}$ , making  $A_t$  measurable by  $\mathcal{F}_t$ . After  $T$  rounds, the decision rule provides the agent’s answer  $\hat{i} \in \{1, 2\}$ , which is measurable by  $\mathcal{F}_T$ . Our objective is to minimize the probability of misidentification, represented as,

$$p_{T} = \mathbb{P}_{\boldsymbol{\mu}} [\hat{i} \neq 1(\boldsymbol{\mu})],$$

where  $1(\boldsymbol{\mu}) = 1\mathbb{1}_{\{\mu_1 > \mu_2\}} + 2\mathbb{1}_{\{\mu_2 > \mu_1\}}$  denotes the best arm under  $\boldsymbol{\mu}$ . Generally, algorithms can adjust the ratio of arm pulls based on the observations. However, an algorithm that operates by pulling arm 1 and arm 2 at a predetermined ratio, regardless of the observational data, and recommends  $\hat{i}$  as the highest empirical average, is referred to as a *static* algorithm. While static algorithms do not encompass all algorithms, they play a pivotal role in this study. Let  $x \in (0, 1)$  denote the portion of the budget allocated to sample the second arm. For a static algorithm parameterized by  $x$ , it pulls the first arm  $(1-x)T + o(T)$  times and the second arm  $xT + o(T)$  times. In particular, an algorithm that samples each arm an equal number of times, where the portion  $x = 1/2$ , is referred to as *uniform sampling*. Define

$$g(x, \boldsymbol{\mu}) = \inf_{\lambda \in (0,1)} (1-x)d(\lambda, \mu_1) + xd(\lambda, \mu_2), \quad (2.7)$$

where  $d(a, b)$  represents the Kullback-Leibler divergence between two Bernoulli distributions with means  $a$  and  $b$  respectively. [30] demonstrated that under the static algorithm parameterized by  $x$ ,

$$\lim_{T \rightarrow \infty} \frac{T}{\log(1/p_{\boldsymbol{\mu}, T})} = \frac{1}{g(x, \boldsymbol{\mu})}.$$

We first introduce a class of *uniformly good* algorithm from [62]:

**Definition 1.** *An algorithm is uniformly good if*

$$\forall \boldsymbol{\mu} \in \Lambda, \quad \limsup_{T \rightarrow \infty} \frac{T}{\log(1/p_{\boldsymbol{\mu}, T})} \leq \frac{1}{g(1/2, \boldsymbol{\mu})}.$$

The uniformly good algorithms are no worse than uniform sampling for all possible bandit instances. The main result of this paper is that any uniformly good algorithm cannot be a strictly better algorithm than uniform sampling. Formally:

**Theorem 7.** *For any uniformly good algorithm,*

$$\forall \boldsymbol{\mu} \in \Lambda, \quad \lim_{T \rightarrow \infty} \frac{T}{\log(1/p_{\boldsymbol{\mu}, T})} = \frac{1}{g(1/2, \boldsymbol{\mu})}.$$

The key to the proof is to show that any uniformly good algorithm is *consistent* and *stable*. They are defined as follows.

**Definition 2.** *An algorithm is consistent if for all  $\boldsymbol{\mu} \in \Lambda$ ,  $\lim_{T \rightarrow \infty} p_{\boldsymbol{\mu}, T} = 0$ .*

**Definition 3.** *An algorithm is stable if for any  $a \in (0, 1)$ , the following properties hold:*

(A) *There exists  $\{\boldsymbol{\lambda}^{(n)}\}_{n=1}^{\infty} \subset \{\boldsymbol{\lambda} \in \Lambda : \lambda_1 > \lambda_2\}$  such that  $\boldsymbol{\lambda}^{(n)} \xrightarrow{n \rightarrow \infty} (a, a)$  and*

$$\lim_{n \rightarrow \infty} \liminf_{T \rightarrow \infty} \mathbb{E}_{\boldsymbol{\lambda}^{(n)}}[\omega_2(T)] = \lim_{n \rightarrow \infty} \limsup_{T \rightarrow \infty} \mathbb{E}_{\boldsymbol{\lambda}^{(n)}}[\omega_2(T)] = \frac{1}{2}.$$

(B) There exists  $\{\boldsymbol{\pi}^{(n)}\}_{n=1}^{\infty} \subset \{\boldsymbol{\pi} \in \Lambda : \pi_1 < \pi_2\}$  such that  $\boldsymbol{\pi}^{(n)} \xrightarrow{n \rightarrow \infty} (a, a)$  and

$$\lim_{n \rightarrow \infty} \liminf_{T \rightarrow \infty} \mathbb{E}_{\boldsymbol{\pi}^{(n)}}[\omega_2(T)] = \lim_{n \rightarrow \infty} \limsup_{T \rightarrow \infty} \mathbb{E}_{\boldsymbol{\pi}^{(n)}}[\omega_2(T)] = \frac{1}{2}.$$

The definition of consistency is adopted from [40]. When the algorithm is stable, it should exhibit continuous behavior (at least in terms of the ratio of arm draws) and demonstrate symmetry across instances. The following theorem may confirm the naturalness of the algorithm classes.

**Theorem 8.** *A uniformly good algorithm is consistent and stable.*

Furthermore, any consistent and stable algorithm has a performance equal to or worse than that of uniform sampling, as stated in the following theorem:

**Theorem 9.** *If an algorithm is consistent and stable, then*

$$\forall \boldsymbol{\mu} \in \Lambda, \quad \liminf_{T \rightarrow \infty} \frac{T}{\log(1/p_{\boldsymbol{\mu}, T})} \geq \frac{1}{g(1/2, \boldsymbol{\mu})}.$$

With Theorems 8 and 9, we are able to prove the main result.

This paper provides partial solutions to the open problems presented in [62].

**Problem 1** in [62] is questioning whether the two following bounds can hold simultaneously:

- *Lower bound.* There exist an algorithm class  $\mathcal{A}$  and a function  $\Gamma^* : \Lambda \mapsto \mathbb{R}$  such that for any algorithm in  $\mathcal{A}$ ,

$$\forall \boldsymbol{\mu} \in \Lambda, \quad \liminf_{T \rightarrow \infty} \frac{T}{\log(1/p_{\boldsymbol{\mu}, T})} \geq \Gamma^*(\boldsymbol{\mu}). \quad (2.8)$$

- *Upper bound.* There is a single algorithm in  $\mathcal{A}$  satisfies

$$\forall \boldsymbol{\mu} \in \Lambda, \quad \limsup_{T \rightarrow \infty} \frac{T}{\log(1/p_{\boldsymbol{\mu}, T})} \leq \Gamma^*(\boldsymbol{\mu}). \quad (2.9)$$

Our Theorem 9 provides a solution to this open problem by defining  $\mathcal{A}$  as the set of consistent and stable algorithms, choosing

$$\Gamma^*(\boldsymbol{\mu}) = \frac{1}{g(1/2, \boldsymbol{\mu})},$$

and the uniform sampling algorithm has the performance matching this lower bound.

**Problem 2** in [62] is questioning the existence of a uniformly good algorithm that strictly outperforms uniform sampling for certain bandit instances. However, according to our Theorem 7, such algorithms do not exist.

**Contribution.** P. Wang and the author of the thesis contributed to the problem formulations through active discussion. P. Wang 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 P. Wang, with all authors actively contributing to subsequent revisions.

## Paper V: Thresholded Lasso Bandit

Paper V is from the following paper.

- Kaito Ariu, Kenshi Abe, and Alexandre Proutiere, “Thresholded Lasso Bandit,” In *Proceedings of the 39th International Conference on Machine Learning (ICML), 2022*.

**Summary.** This paper considers a high-dimensional contextual linear stochastic bandit problem. At each round, denoted as  $t \in [T] = \{1, \dots, T\}$ , the algorithm receives a set of context vectors  $\mathcal{A}_t = \{A_{t,k} \in \mathbb{R}^d : k \in [K]\}$ . The sets  $(\mathcal{A}_t)_{t \geq 1}$  form an independent and identically distributed random matrix sequence with a distribution  $p_A$ . For each round  $t$ , the algorithm sequentially chooses an arm  $A_t \in \mathcal{A}_t$  and receives a random reward  $r_t$ . Let  $\mathcal{F}_t$  be the  $\sigma$ -algebra generated by the random variables  $(\mathcal{A}_1, A_1, r_1, \dots, \mathcal{A}_{t-1}, A_{t-1}, r_{t-1}, \mathcal{A}_t)$ ,  $A_t$  is  $\mathcal{F}_t$ -measurable. The reward  $r_t$  is assumed to be the inner product of  $A_t$  and  $\theta$ , plus a sub-Gaussian random variable  $\varepsilon_t$  with variance proxy  $\sigma^2$  given  $\mathcal{F}_t$  and  $A_t$ . We make the assumption that  $\theta$  has at most  $s_0 \ll d$  non-zero elements. The decision-maker is unaware of the set of non-zero elements and its cardinality  $s_0$ . The objective of the study is to design an algorithm that minimizes regret, which is defined as follows:

$$\begin{aligned} R(T) &= \mathbb{E} \left[ \sum_{t=1}^T \max_{A \in \mathcal{A}_t} \langle A, \theta \rangle - r_t \right] \\ &= \mathbb{E} \left[ \sum_{t=1}^T \max_{A \in \mathcal{A}_t} \langle A - A_t, \theta \rangle \right]. \end{aligned}$$

We use the following notation. The  $\ell_0$ -norm of a vector  $\theta$  in  $\mathbb{R}^d$  is represented as  $\|\theta\|_0 = \sum_{i=1}^d \mathbb{1}\{\theta_i \neq 0\}$ , essentially summing up the non-zero elements in the vector. We use  $\hat{\Sigma}_t = \frac{1}{t} \sum_{s=1}^t A_s A_s^\top$  to denote the empirical Gram matrix, which is constructed based on the arms chosen by a certain algorithm. For any  $B \subset [d]$ , we define  $\theta_B$  as  $(\theta_{1,B}, \dots, \theta_{d,B})^\top$ , where for each  $i$  in  $[d]$ ,  $\theta_{i,B} = \theta_i \mathbb{1}\{i \in B\}$ . We also define a submatrix  $A(B) \in \mathbb{R}^{n \times |B|}$  for each subset  $B$  within  $[d]$ . This submatrix is part of the matrix  $A$  in  $\mathbb{R}^{n \times d}$ , and for  $A(B)$ , we only include the rows that exist in  $B$ . The symbol  $\text{supp}(x)$  represents the set of indices of non-zero elements of  $x$  in  $\mathbb{R}^d$ . We also define  $\theta_{\min}$  as the smallest absolute value of  $\theta_i$  on the support, i.e.,  $\theta_{\min} = \min_{i \in \text{supp}(\theta)} |\theta_i|$ , where the support of  $\theta$  is denoted by  $S(\theta) = \text{supp}(\theta) = \{i \in [d] : \theta_i \neq 0\}$ .

The paper introduces the following assumptions. Some are adopted from [60].

**Assumption 2** (Parameters). *The reward function is defined by the parameter  $\theta$ , which is sparse. This means that  $\|\theta\|_0 \leq s_0$  for some unknown  $s_0$  ( $s_0$  does not depend on  $d$ ). Additionally, we assume that  $\|\theta\|_1 \leq s_1$ , and  $\theta_{\min} \geq s_2/s_0$ . Lastly, we assume that the infinity-norm of the context vector is bounded:  $\forall t, \forall A \in \mathcal{A}_t, \|A\|_\infty \leq s_A$ .*

**Assumption 3** (Compatibility condition). *For each  $M \in \mathbb{R}^{d \times d}$  and  $S_0 \subset [d]$ , the compatibility constant  $\phi(M, S_0)$  is defined as:*

$$\phi^2(M, S_0) = \min_{x: \|x_{S_0}\|_1 \neq 0} \left\{ \frac{s_0 x^\top M x}{\|x_{S_0}\|_1^2} : \|x_{S_0^c}\|_1 \leq 3\|x_{S_0}\|_1 \right\}.$$

We assume that the expected gram matrix for the actions  $\Sigma = \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{A \sim p_A} [A_k A_k^\top]$  has the following property:

$$\phi^2(\Sigma, S(\theta)) \geq \phi_0^2,$$

where  $\phi_0 > 0$  is some constant.

**Assumption 4** (Relaxed symmetry [60]). *For the distribution  $p_A$  of  $\mathcal{A}$ , there exists a constant  $\nu \geq 1$  such that for all  $\mathbf{A} \in \mathbb{R}^{K \times d}$  such that  $p_A(\mathbf{A}) > 0$ ,  $\frac{p_A(\mathbf{A})}{p_A(-\mathbf{A})} \leq \nu$ .*

**Assumption 5** (Balanced covariance [60]). *For any permutation  $\gamma$  of  $[K]$ , for any integer  $k \in \{2, \dots, K-1\}$  and a fixed  $\theta$ , there exists a constant  $C_b > 1$  such that*

$$\begin{aligned} C_b \mathbb{E}_{A \sim p_A} & \left[ (A_{\gamma(1)} A_{\gamma(1)}^\top + A_{\gamma(K)} A_{\gamma(K)}^\top) \right. \\ & \left. \cdot \mathbf{1}\{\langle A_{\gamma(1)}, \theta \rangle < \dots < \langle A_{\gamma(K)}, \theta \rangle\} \right] \\ & \succeq \mathbb{E}_{A \sim p_A} \left[ A_{\gamma(k)} A_{\gamma(k)}^\top \mathbf{1}\{\langle A_{\gamma(1)}, \theta \rangle < \dots < \langle A_{\gamma(K)}, \theta \rangle\} \right]. \end{aligned}$$

**Assumption 6** (Sparse positive definiteness). *Define for each  $B \subset [d]$   $\Sigma_B = \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{A \sim p_A} [A_k(B) A_k(B)^\top]$ , where  $A_k(B) \in \mathbb{R}^{|B|}$  is a subvector from the elements of  $A_k$  with indices in  $B$ . There exists a constant  $\alpha > 0$  such that for all  $B \subset [d]$ ,*

$$\begin{aligned} |B| \leq s_0 + (4\nu C_b \sqrt{s_0}) / \phi_0^2 \text{ and } S(\theta) \subset B \\ \implies \min_{v \in \mathbb{R}^{|B|}: \|v\|_2=1} v^\top \Sigma_B v \geq \alpha. \end{aligned}$$

The parameters  $\phi_0, \nu, C_b$  are those of Assumptions 3, 4, and 5.

Some theorems use the following margin condition, which controls the probability that two arms yield very similar rewards. It is widely discussed in the classification literature [8, 72].

**Assumption 7** (Margin condition). *There exists a constant  $C_m > 0$  such that for all  $\kappa > 0$ ,*

$$\forall k \neq k', \quad \mathbb{P}_{A \sim p_A}(0 < |\langle A_k - A_{k'}, \theta \rangle| \leq \kappa) \leq C_m \kappa.$$

Under this setting, we propose a Thresholded Lasso bandit (TH Lasso bandit) that sequentially pulls the arm greedily and estimates  $\theta$  and its support. We provide the following non-asymptotic regret upper bound of TH Lasso bandit with the margin condition. Let  $\tau = \left\lfloor \frac{2 \log(2d^2)}{C_0^2} (\log s_0) (\log \log d) \right\rfloor$ , where  $C_0 = \min \left\{ \frac{1}{2}, \frac{\phi_0^2}{512 s_0 s_A^2 \nu C_b} \right\}$ . We remark that  $\tau = \mathcal{O}(s_0^2 (\log s_0) (\log d) (\log \log d))$ .

**Theorem 10.** *Assume that Assumptions 2 – 6, 7 hold.*

(i) *(TH Lasso Bandit with parameter-dependent input) There exist constants  $c_1, c_2, c_3 > 0$  depending on  $\sigma, s_A, s_1, s_2, \phi_0, \nu, C_b, K, \alpha, C_m$ , such that if we set  $\lambda_0 = c_1$ , for all  $d \geq c_2$ , for all  $T \geq 2$ :*

$$R(T) \leq c_3 \left( \tau + s_0 (\log s_0)^{\frac{3}{2}} \log T + s_0^2 \right).$$

(ii) *(TH Lasso Bandit with parameter-free input) There exist constants  $c_4, c_5 > 0$  depending on  $\sigma, s_A, s_1, s_2, \phi_0, \nu, C_b, K, \alpha, C_m$ , such that if we set  $\lambda_0 = 1/(\log \log d)^{\frac{1}{4}}$ , for all  $d \geq c_4$ , for all  $T \geq 2$ ,*

$$R(T) \leq c_5 \left( \tau + s_0 (\log s_0)^{\frac{3}{2}} \log T + s_0^2 \right).$$

*The definitions of  $c_1$ - $c_5$  are given in Appendix of [2].*

We also provide the following non-asymptotic regret upper bound of TH Lasso bandit without the margin condition.

**Theorem 11.** *Assume that Assumptions 2 – 6 hold.*

(i) *(TH Lasso Bandit with parameter-dependent input) There exist constants  $c_1, c_2, c_3 > 0$  depending on  $\sigma, s_A, s_1, s_2, \phi_0, \nu, C_b, K, \alpha$  such that if we set  $\lambda_0 = c_1$ , for all  $d \geq c_2$ , for all  $T \geq 2$ :*

$$R(T) \leq c_3 \left( \tau + (\log s_0) \sqrt{s_0 T} + s_0^2 \right).$$

(ii) *(TH Lasso Bandit with parameter-free input) There exist constants  $c_4, c_5 > 0$  depending on  $\sigma, s_A, s_1, s_2, \phi_0, \nu, C_b, K, \alpha$  such that if we set  $\lambda_0 = 1/(\log \log d)^{\frac{1}{4}}$  in TH Lasso Bandit, for all  $d \geq c_4$ , for all  $T \geq 2$ ,*

$$R(T) \leq c_5 \left( \tau + (\log s_0) \sqrt{s_0 T} + s_0^2 \right).$$

*The definitions of  $c_1$ - $c_5$  are given in Appendix of [2].*

Theorems 10 and 11 achieve much lower regret than the existing algorithms. Their scalings are  $\log d \log T$  (resp.  $\log d + \sqrt{T \log(dT)}$ ) with (resp. without) the margin condition. These guarantees are much lower compared with existing guarantees [10, 42, 60, 75]. Moreover, we have numerically confirmed the superiority of our algorithm for various problem instances.

**Contribution.** The author of the thesis and K. Abe introduced the model and problem formulation through active discussion. The author of the thesis designed the algorithm and established majority of the proof. K. Abe and A. Proutiere provided some keys in the proof. K. Abe conducted numerical experiments with synthetic and real-world data. The author of the thesis wrote the majority of the manuscript, with all authors actively contributing to the revision of the draft.

## Paper VI: Regret in Online Recommendation System

Paper VI is from the following paper.

- Kaito Ariu, Narae Ryu, Se-Young Yun, and Alexandre Proutiere, “Regret in Online Recommendation Systems,” In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

**Summary.** This paper conducts theoretical analyses of online recommendation systems. We consider a system that includes a collection of items, labeled as  $\mathcal{I} = [n] = \{1, \dots, n\}$ , and a group of users, labeled as  $\mathcal{U} = [m]$ . For each round, a randomly chosen user from  $\mathcal{U}$  requires a recommendation. The decision-maker, after observing the user’s identity, recommends an item to the user. An important assumption is that a user cannot receive the same recommendation twice, or doing so yields no reward. The user immediately provides a rating for the recommended item, giving it a +1 if they like it or a 0 if they do not. The decision-maker then observes this rating immediately. Specifically, in round  $t \in [T]$ , the user  $u_t$  is chosen randomly from  $\mathcal{U}$  and needs a recommendation. If item  $i$  is recommended, the user  $u_t = u$  will like the item with a probability of  $\rho_{iu}$ . We use a random variable  $X_{iu}$  to represent whether the user likes the item.  $X_{iu}$  has a Bernoulli distribution with parameter  $\rho_{iu}$ . Let  $\pi$  represent a sequential item selection strategy or algorithm. Under the algorithm  $\pi$ , the item  $i_t^\pi$  is presented to the  $t$ -th user. The choice  $i_t^\pi$  is dependent on past observations and the identity of the  $t$ -th user, meaning  $i_t^\pi$  is measurable with respect to  $\mathcal{F}_{t-1}^\pi$  where  $\mathcal{F}_{t-1}^\pi = \sigma(u_t, (u_s, i_s^\pi, X_{i_s^\pi u_s}), s \leq t-1)$ . The reward of an algorithm  $\pi$  is defined as the expected number of positive ratings received over  $T$  rounds:  $\mathbb{E}[\sum_{t=1}^T \rho_{i_t^\pi u_t}]$ . Our goal is to develop an algorithm that maximizes this reward. Our primary focus is on scenarios where the variables  $(m, n, T)$  grows large while adhering to the constraints (i)  $m \geq n$  (which is usually the case in recommendation systems), (ii)  $T = o(mn)$ , and (iii)  $\log(m) = o(n)$ .

Let  $N_u(T)$  represents the total number of recommendation for user  $u$  up to round  $T$ . We use the following fact from the literature on *Balls and Bins process*

[63]: define  $\bar{n} = \mathbb{E}[\max_{u \in \mathcal{U}} N_u(T)]$ , then

$$\bar{n} = \begin{cases} \frac{\log(m)}{\log(\frac{m \log(m)}{T})} (1 + o(1)) & \text{if } T = o(m \log(m)), \\ \log(m)(d_c + o(1)) & \text{if } T = cm \log(m), \\ \frac{T}{m} (1 + o(1)) & \text{if } T = \omega(m \log(m)), \end{cases}$$

where  $c$  and  $d_c$  are positive constants. We also obtain a following tail bound for  $N_u(T)$ .

**Lemma 1.** *Define*

$$\bar{N} = \frac{4 \log(m)}{\log(\frac{m \log(m)}{T} + e)} + \frac{e^2 T}{m}.$$

*Then,*

$$\forall u \in \mathcal{U}, \quad \mathbb{E}[\max\{0, N_u(T) - \bar{N}\}] \leq \frac{1}{(e-1)m}.$$

For the success rates  $\rho = (\rho_{iu})_{i \in \mathcal{I}, u \in \mathcal{U}}$ , three different types of structural assumptions are made. For each model, regret lower bound and upper bound are provided.

*Model A. Clustered items and statistically identical users.* In this model,  $\rho_{iu}$  depends only on the item index  $i$ . The items are partitioned into  $K$  clusters  $\mathcal{I}_1, \dots, \mathcal{I}_K$ . When an item  $i$  is recommended for the first time, it is assigned to cluster  $\mathcal{I}_k$  with a probability of  $\alpha_k$ , independent of the cluster assignments of other items. If  $i \in \mathcal{I}_k$ , then  $\rho_i = p_k$ . We assume that both  $\alpha = (\alpha_k)_{k \in [K]}$  and  $p = (p_k)_{k \in [K]}$  are initially unknown and do not depend on  $(n, m, T)$ . Without loss of generality, assume that  $p_1 > p_2 \geq p_3 \geq \dots \geq p_K$ . The regret of an algorithm  $\pi \in \Pi$  is defined as the difference between its reward and that of an Oracle algorithm, which is aware of the item clusters and the parameters  $p$ . We define the regret as if recommending items from  $\mathcal{I}_1$  was always possible. The regret of  $\pi \in \Pi$  is:  $R^\pi(T) = Tp_1 - \sum_{t=1}^T \mathbb{E} \left[ \sum_{k=1}^K \mathbb{1}_{\{i_t^\pi \in \mathcal{I}_k\}} p_k \right]$ . Define  $\Delta_k = p_1 - p_k$  as the difference in success rates between the best-performing item cluster and the items in cluster  $\mathcal{I}_k$ . We then define the function:

$$\phi(k, m, p) = \frac{1 - e^{-m\gamma(p_1, p_k)}}{8(1 - e^{-\gamma(p_1, p_k)})},$$

where  $\gamma(p, q) = \text{KL}(p, q) + \text{KL}(q, p)$  and  $\text{KL}(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ . Given that  $\text{KL}(p, q) \leq (p-q)^2 / (q(1-q))$ , it can be demonstrated that as  $m$  grows large,  $\phi(k, m, p)$  scales as  $\eta / (16\Delta_k^2)$  for small  $\Delta_k$ , with  $\eta = \min_k p_k(1-p_k)$ . We say an algorithm  $\pi$  is uniformly good if the following hold: for any  $(p, \alpha)$ ,  $R^\pi(T)$  is  $O(\max\{\sqrt{T}, \frac{\log(m)}{\log(\frac{m \log(m)}{T} + e)}\})$  as  $T, m, n$  grow large with  $T = o(nm)$  and  $m \geq n$ .

We obtained the following regret lower bounds.

**Theorem 12** (Regret lower bounds for Model A). *Let  $\pi \in \Pi$  be any algorithm. For all  $T \geq 1$  such that  $m \geq c/\Delta_2^2$  ( $c$  is some positive constant),*

$$R^\pi(T) \geq \max\{R_{\text{nr}}(T), R_{\text{ic}}(T)\},$$

where  $R_{\text{nr}}(T)$  represents the regret due to the no-repetition constraint and  $R_{\text{ic}}(T)$  represents the regret due to the unknown item clusters. They are defined as follows:

$$R_{\text{nr}}(T) = \bar{n} \sum_{k \neq 1} \alpha_k \Delta_k$$

and

$$R_{\text{ic}}(T) = \frac{T}{m} \sum_{k \neq 1} \alpha_k \phi(k, m, p) \Delta_k.$$

Furthermore, for any uniformly good algorithm  $\pi$ ,

$$R^\pi(T) \gtrsim R_{\text{sp}}(T) = \log(T) \sum_{k \neq 1} \frac{\Delta_k}{2\text{KL}(p_k, p_1)},$$

where  $R_{\text{sp}}(T)$  represents the regret due to the unknown success probabilities and we write  $a \gtrsim b$  if  $\liminf_{T \rightarrow \infty} a/b \geq 1$ .

We propose ECT, an algorithm that is composed of three phases: (i) Exploration, (ii) Clustering, and (iii) Test. In the exploration phase, samples are gathered from a subset of randomly selected items to learn their success probabilities and clusters. The clustering phase uses the gathered information to estimate each item's success probability and clusters the items using a variant of the K-means algorithm. Lastly, in the test phase, items believed to be in the best cluster are recommended, their success rates are continually updated, and items that have less confidence to be in the best cluster are removed. ECT works without breaking the no-repetition constraint for all of the phases. The following regret guarantee is obtained.

**Theorem 13** (Regret upper bound for Model A). *We have,*

$$R^{\text{ECT}}(T) = \mathcal{O}\left(\frac{2\bar{N}}{\alpha_1} \sum_{k=2}^K \frac{\alpha_k(p_1 - p_k)}{(p_1 - p_2)^2} + (\log T)^3\right).$$

According to Theorem 12, the regret of any algorithm  $\pi$  is at least  $\Omega(\bar{N})$ , and particularly if  $\pi$  is uniformly good, regret is lower bounded as  $\Omega(\max\{\bar{N}, \log(T)\})$ . From Theorem 13, if  $\bar{N}$  is  $\Omega((\log T)^3)$ , ECT exhibits optimal order, and otherwise, it exhibits optimal order up to a factor of  $(\log T)^2$ . Furthermore, when  $R_{\text{ic}}(T)$  ( $= \Omega(\frac{T}{\Delta_2 m})$ ) is the largest part of the regret lower bound, the regret of ECT also scales with  $\Delta_2$ :  $R^{\text{ECT}}(T) = \mathcal{O}(\frac{T}{\Delta_2 m})$ .

*Model B. Unclustered items and statistically identical users.* In this model,  $\rho_{iu}$  depends only on the item  $i$ . When a new item  $i$  is recommended for the first time, its success rate  $\rho_i$  is drawn from some distribution  $\zeta$  over  $[0, 1]$ , independent of the success rates of other items.  $\zeta$  is initially unknown and arbitrary, but for simplicity, we assume it to be absolutely continuous with respect to the Lebesgue measure. We adopt the following notion of *satisficing regret* [65]: for a given  $\varepsilon > 0$ ,

$$R_\varepsilon^\pi(T) = \sum_{t=1}^T \mathbb{E} [\max\{0, \mu_{1-\varepsilon} - \rho_{i_t^\pi}\}].$$

Recommending items within the  $\varepsilon$ -best items does not generate any satisficing regret, and it is assumed that an Oracle policy can always recommend such items. The satisficing regret is caused by two factors: the no-repetition constraint and learning the item's success rate.

**Theorem 14** (Regret lower bounds for Model B). *Assume  $\zeta(\mu) \leq C$  for all  $\mu \in [0, 1]$  with some positive constant  $C$ . Let  $\pi \in \Pi$  be an arbitrary algorithm. For all  $T \geq 1$  such that  $m \geq c/\varepsilon^2$  (for some constant  $c \geq 1$ ),*

$$R_\varepsilon^\pi(T) \geq \max\{R_{\text{nr}}(T), R_i(T)\},$$

where

$$R_{\text{nr}}(T) = \bar{n} \int_0^{\mu_{1-\varepsilon}} (\mu_{1-\varepsilon} - \mu) \zeta(\mu) d\mu$$

and

$$R_i(T) = \frac{T}{m} \frac{\frac{(1-\varepsilon)^2}{2C} \left(1 - \frac{\varepsilon C}{1-\varepsilon}\right)^2}{\min\{1, (1+C)\varepsilon\} + 1/m}.$$

We propose ET (Explore-and-Test) algorithm with two phases: (i) an exploration phase that aims at estimating the threshold level  $\mu_{1-\varepsilon}$  and (ii) a test phase where we apply to each item sequential tests to determine whether the item is above the threshold. The following satisficing regret bound is obtained.

**Theorem 15** (Regret upper bound for Model B). *Assume that the following condition is satisfied:  $\zeta(\mu) \leq C$  for all  $\mu \in [0, 1]$ .*

*For any  $\varepsilon \geq C \sqrt{\frac{\pi}{2 \log T}}$ , the satisficing regret of ET is bounded as follows.*

$$R_\varepsilon^{\text{ET}}(T) = \mathcal{O} \left( \bar{N} \frac{\log(1/\varepsilon) \log \log(m)}{\varepsilon} + \frac{(\log T)^2}{\varepsilon^2} \right).$$

When we take into account Theorem 14, the satisficing regret of any algorithm scales at least  $\Omega(\frac{\bar{N}}{\varepsilon})$ . Consequently, as per Theorem 15, ET achieves the optimality, at least when  $\bar{N} = \Omega((\log T)^2)$ .

*Model C. Clustered items and clustered users.* In this model, both items and users are clustered. Users are grouped into  $L$  clusters  $\mathcal{U}_1, \dots, \mathcal{U}_L$ , and when a user first arrives at the system, she is assigned to cluster  $\mathcal{U}_\ell$  with probability  $\beta_\ell$ , independent of other users. There are  $K$  item clusters  $\mathcal{I}_1, \dots, \mathcal{I}_K$ . When the algorithm recommends an item  $i$  for the first time, it is assigned to cluster  $\mathcal{I}_k$  with probability  $\alpha_k$  as in Model A. Now,  $\rho_{iu} = p_{k\ell}$  when  $i \in \mathcal{I}_k$  and  $u \in \mathcal{U}_\ell$ . We assume that an Oracle algorithm, aware of the item and user clusters and of the parameters  $p$ , would only recommend items from cluster  $k_\ell^*$  to a user in  $\mathcal{U}_\ell$ . The regret of an algorithm  $\pi \in \Pi$  is defined as:

$$R^\pi(T) = T \sum_{\ell} \beta_\ell p_{k_\ell^* \ell} - \sum_{t=1}^T \mathbb{E} \left[ \sum_{k, \ell} \mathbf{1}_{\{u_t \in \mathcal{U}_\ell, i_t^\pi \in \mathcal{I}_k\}} p_{k\ell} \right].$$

We introduce the following notations. For any  $\ell \in [L]$ , let  $\Delta_{k\ell} = p_{k_\ell^* \ell} - p_{k\ell}$  denote the gap between the success rates of items from the top cluster  $\mathcal{I}_{k_\ell^*}$  and items from cluster  $\mathcal{I}_k$ . We introduce  $\mathcal{R}_\ell = \{r \in [L] : k_\ell^* \neq k_r^*\}$  to denote the set of user clusters that have different best items. Define  $\mathcal{L}_\perp = \{(\ell, \ell') \in [L]^2 : p_{k_\ell^* \ell} \neq p_{k_{\ell'}^* \ell'}\}$ , the set of pairs of user clusters where the best item clusters differ. Additionally, we introduce the functions:

$$\phi(k, \ell, m, p) = \frac{1 - e^{-m\gamma(p_{k_\ell^* \ell}, p_{k\ell})}}{8 \left(1 - e^{-\gamma(p_{k_\ell^* \ell}, p_{k\ell})}\right)}$$

and

$$\psi(\ell, k, T, m, p) = \frac{1 - e^{-\frac{T}{m}\gamma(p_{k_\ell^* \ell}, p_{k\ell})}}{8 \left(1 - e^{-\gamma(p_{k_\ell^* \ell}, p_{k\ell})}\right)}.$$

In contrast to Model A, the algorithm can exploit the user clusters. If  $\mathcal{L}_\perp \neq \emptyset$ , then there are multiple optimal item clusters depending on the users, and when a user  $u$  first arrives, it is necessary to learn its cluster. This learning of the user cluster induces at least a constant regret per user. We define the regret component induced by learning the user cluster as follows.

$$R_{\text{uc}}(T) = m \sum_{\ell \in [L]} \beta_\ell \frac{\sum_{k \in \mathcal{R}_\ell} \Delta_{k\ell} \psi(\ell, k, T, m, p)}{K}.$$

For specific values of  $p$ , we illustrate that this classification can even generate a regret that scales as  $\log(T/m)$  (per user). This occurs when  $\mathcal{L}_\perp^\perp(\ell) = \{\ell' \neq \ell : k_\ell^* \neq k_{\ell'}^*, p_{k_\ell^* \ell} = p_{k_{\ell'}^* \ell'}\}$  is not empty. In this case, it is impossible to distinguish users from  $\mathcal{U}_\ell$  and  $\mathcal{U}_{\ell'}$  by merely presenting items from  $\mathcal{I}_{k_\ell^*}$  (the greedy choice for users in  $\mathcal{U}_\ell$ ). The corresponding regret term is defined as follows.

$$R'_{\text{uc}}(T) = c(\beta, p)m \log(T/m)$$

where

$$c(\beta, p) = \inf_{n \in \mathcal{F}} \sum_{\ell} \beta_{\ell} \sum_{k \neq k_{\ell}^*} \Delta_{k\ell} n_{k\ell}$$

with

$$\mathcal{F} = \{n \geq 0 : \forall \ell, \forall \ell' \in \mathcal{L}^{\perp}(\ell), \sum_{k \neq k_{\ell}^*} \text{KL}(p_{k\ell}, p_{k\ell'}) n_{k\ell} \geq 1\}.$$

The component of regret from the no-repetition constraint is defined similarly as in Model A:

$$R_{\text{nr}}(T) = \bar{n} \sum_{\ell} \beta_{\ell} \sum_{k \neq k_{\ell}^*} \alpha_k \Delta_{k\ell}.$$

Finally, an algorithm is uniformly good if for any user  $u$ ,  $R_u^{\pi}(N) = o(N^{\alpha})$  as  $N$  grows large for all  $\alpha > 0$ , where  $R_u^{\pi}(N)$  represents regret under  $\pi$  for user  $u$  when the user  $u$  has arrived  $N$  times. We obtain the following regret lower bounds.

**Theorem 16** (Regret lower bounds for Model C). *For any algorithm  $\pi \in \Pi$ , for all  $T \geq 2m$  such that  $m \geq c / \min_{k,\ell} \Delta_{k\ell}^2$  with for some constant  $c > 0$ ,*

$$R^{\pi}(T) \geq \max\{R_{\text{nr}}(T), R_{\text{ic}}(T), R_{\text{uc}}(T)\}.$$

*Additionally, when  $T = \omega(m)$ , if  $\pi$  is uniformly good,*

$$R^{\pi}(T) \gtrsim R'_{\text{uc}}(T) = c(\beta, p)m \log(T/m),$$

where

$$c(\beta, p) = \inf_{n \in \mathcal{F}} \sum_{\ell} \beta_{\ell} \sum_{k \neq k_{\ell}^*} \Delta_{k\ell} n_{k\ell}$$

with

$$\mathcal{F} = \{n \geq 0 : \forall \ell, \forall \ell' \in \mathcal{L}^{\perp}(\ell), \sum_{k \neq k_{\ell}^*} \text{KL}(p_{k\ell}, p_{k\ell'}) n_{k\ell} \geq 1\}.$$

When designing the algorithm, we need to take into account the inability to control the user arrival process. Consequently, it is not straight forward to cluster users as what we did for clustering the items. We propose an algorithm called Explore-Cluster with Upper Confidence Sets (EC-UCS). The concept behind EC-UCS is: we estimate the success rates  $(p_{k\ell})_{k,\ell}$  by using small subsets of items and users. Then, based on these estimates, each user is optimistically assigned to an Upper Confidence Set (UCS), a set of clusters that the user is likely to belong to. As the number of requests made by a user increases, the user's UCS shrinks (similar to how the UCB index of an arm in bandit problems approaches its average reward). This algorithm is composed of three main phases. The first

phase involves collecting data to infer item clusters, and creating clusters based on user responses to a randomly selected set of items. The second phase clusters users, extracting clusters from a subset of the most recommended users. The final phase makes recommendations based on estimated success probabilities, and in cases where the user's cluster is uncertain, exploration is conducted to select the best cluster optimistically. Our analysis yields the following regret bound for EC-UCS.

**Theorem 17** (Regret upper bound for Model C). *For any  $\ell$ , let  $\sigma_\ell$  be the permutation of  $[K]$  such that  $p_{\sigma_\ell(1)\ell} > p_{\sigma_\ell(2)\ell} \geq \dots \geq p_{\sigma_\ell(K)\ell}$ . Let  $\mathcal{S}_{\ell r} = \{k \in [K] : p_{k\ell} \neq p_{kr}\}$ ,  $y_{\ell r} = \min_{k \in \mathcal{S}_{\ell r}} |p_{k\ell} - p_{kr}|$ ,  $\delta = \min_\ell (p_{\sigma_\ell(1)\ell} - p_{\sigma_\ell(2)\ell})$ , and  $\phi(x) = x/\log(1/x)$ . Then, the following holds:*

$$R^{\text{EC-UCS}}(T) = \mathcal{O} \left( m \sum_{\ell} \beta_{\ell} (p_{\sigma_{\ell}(1)\ell} - p_{\sigma_{\ell}(K)\ell}) \left( \max \left( \frac{K^3 \log K}{\phi(\min(y_{\ell r}, \delta)^2)}, \frac{\sqrt{K}}{\min_{\ell} \beta_{\ell}} \right) + \sum_{r \in \mathcal{R}_{\ell} \setminus \mathcal{L}^{\perp}(\ell)} \frac{K^2 \log K}{\phi(|p_{k_{\ell}^* r} - p_{k_{\ell}^* \ell}|^2)} + \sum_{k \in \mathcal{S}_{\ell r}} \sum_{r \in \mathcal{L}^{\perp}(\ell)} \frac{K \log \bar{N}}{|\mathcal{S}_{\ell r}| |p_{k\ell} - p_{kr}|^2} \right) \right).$$

The EC-UCS algorithm's regret aligns with our lower bound in terms of order. Specifically, the algorithm achieves a regret that (i) scales with  $m$  whenever feasible, that is, when  $\mathcal{L}^{\perp}(\ell) = \emptyset$  for every  $\ell$ , and (ii) scales with  $m \log(\bar{N})$  otherwise.

**Contribution.** The author of the thesis formulated the problem through active discussion with A. Proutiere. The author of the thesis, together with N. Ryu, and S. Yun established the theoretical results. The author of the thesis also conducted numerical experiments. All authors contributed to the writing and revision of the 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] Kaito Ariu, Kenshi Abe, and Alexandre Proutière. Thresholded lasso bandit. In *International Conference on Machine Learning*, pages 878–928. PMLR, 2022.
- [3] 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.
- [4] Kaito Ariu, Jungseul Ok, Alexandre Proutiere, and Se-Young Yun. Optimal clustering from noisy binary feedback. *arXiv preprint arXiv:1910.06002*, 2019.
- [5] Kaito Ariu, Alexandre Proutiere, and Se-Young Yun. Instance-optimal cluster recovery in the labeled stochastic block model. *arXiv preprint arXiv:2306.12968*, 2023.
- [6] Kaito Ariu, Narae Ryu, Se-Young Yun, and Alexandre Proutière. Regret in online recommendation systems. *Advances in Neural Information Processing Systems*, 33:21141–21150, 2020.
- [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] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 2007.
- [9] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding clinical trials. *The Journal of Machine Learning Research*, 22(1):686–723, 2021.
- [10] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

- [11] Thomas Bonald and Alexandre Proutiere. Two-target algorithms for infinite-armed bandits with bernoulli rewards. In *Advances in Neural Information Processing Systems 26*, pages 2184–2192. 2013.
- [12] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International conference on machine learning*, pages 793–802. PMLR, 2019.
- [13] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.
- [14] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- [15] Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.
- [16] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.
- [17] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.
- [18] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.
- [19] Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. *Advances in Neural Information Processing Systems*, 30, 2017.
- [20] Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529. PMLR, 2014.
- [21] Rémy Degenne. On the existence of a complexity in fixed budget bandit identification. *arXiv preprint arXiv:2303.09468*, 2023.
- [22] Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.
- [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] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.
- [25] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.
- [26] Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 88–97, 1994.
- [27] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.
- [28] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024, 2017.
- [29] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.
- [30] 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.
- [31] Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.
- [32] Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.
- [33] Simon Heimlicher, Marc Lelarge, and Laurent Massoulié. Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*, 2012.
- [34] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
- [35] Yassir Jedra, Junghyun Lee, Alexandre Proutiere, and Se-Young Yun. Nearly optimal latent state decoding in block MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 2805–2904. PMLR, 2023.

- [36] Yassir Jedra and Alexandre Proutiere. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control*, pages 2676–2681. IEEE, 2019.
- [37] Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.
- [38] 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.
- [39] Emilie Kaufmann. Contributions to the optimal solution of several bandits problems, 2020.
- [40] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of A/B testing. In *Conference on Learning Theory*, pages 461–481. PMLR, 2014.
- [41] 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.
- [42] Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, 2019.
- [43] Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin. Rate-optimal bayesian simple regret in best arm identification. *Mathematics of Operations Research*, 2023.
- [44] 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.
- [45] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *The Journal of Machine Learning Research*, 22(1):1395–1476, 2021.
- [46] Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics*, pages 1091–1114, 1987.
- [47] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- [48] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

- [49] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- [50] Marc Lelarge, Laurent Massoulié, and Jiaming Xu. Reconstruction in the labeled stochastic block model. In *2013 IEEE Information Theory Workshop*, 2013.
- [51] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.
- [52] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016.
- [53] 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.
- [54] Jérémie Mary, Romaric Gaudel, and Philippe Preux. Bandits and recommender systems. In *International Workshop on Machine Learning, Optimization and Big Data*, pages 325–336, 2015.
- [55] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.
- [56] Tetsuro Morimura. *Reinforcement learning (in Japanese)*. Kodansha, 2019.
- [57] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.
- [58] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.
- [59] Thanh T Nguyen, Charles Soussen, Jérôme Idier, and El-Hadi Djermoune. NP-hardness of  $\ell_0$  minimization problems: revision and extension to the non-negative setting. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–4. IEEE, 2019.
- [60] Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, 2021.
- [61] Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

- [62] Chao Qin. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, pages 5650–5654. PMLR, 2022.
- [63] Martin Raab and Angelika Steger. Balls into Bins - A Simple and Tight Analysis. In *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 159–170, 1998.
- [64] Lior Rokach and Oded Maimon. Clustering methods. *Data mining and knowledge discovery handbook*, pages 321–352, 2005.
- [65] Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *arXiv preprint arXiv:1803.02855*, 2018.
- [66] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [67] Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.
- [68] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [69] 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.
- [70] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [71] Andrea Tirinzoni, Alessandro Lazaric, and Marcello Restelli. A novel confidence-based algorithm for structured bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3175–3185. PMLR, 2020.
- [72] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 2004.
- [73] 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.
- [74] 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.
- [75] Xue Wang, Mingcheng Wei, and Tao Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, 2018.

- [76] Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.
- [77] Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.
- [78] Se-Young Yun and Alexandre Proutière. Optimal sampling and clustering in the stochastic block model. In *Advances in Neural Information Processing Systems*, pages 13422–13430, 2019.
- [79] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022.
- [80] Shuheng Zhou. Thresholded lasso for high dimensional variable selection and statistical estimation, 2010.



# Appended Papers

