Thesis for The Degree of Doctor of Philosophy

Leveraging Structural Priors and Historical Data for Practical Treatment Personalization with Multi-Armed Bandits

Newton Mwai

Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2025

Leveraging Structural Priors and Historical Data for Practical Treatment Personalization with Multi-Armed Bandits

Newton Mwai

© Newton Mwai, 2025
except where otherwise stated.
All rights reserved.

ISBN 978-91-8103-295-6
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5753.
ISSN 0346-718X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2025.

To my family.

Leveraging Structural Priors and Historical Data for Practical Treatment Personalization with Multi-Armed Bandits

Newton Mwai

Department of Computer Science and Engineering
Chalmers University of Technology

Abstract

Personalizing treatments for patients often requires sequentially trying different options from a set of available therapies until the most effective one is identified for the patient's characteristics. In chronic diseases such as Alzheimer's Disease, where interventions mainly have short-term effects, this search process can be formulated as a multi-armed bandit (MAB) problem. Reducing the length of the search is essential to limit patient burden and other associated costs, while practical constraints, such as limiting switches between therapies, introduce additional complexity to exploration. This thesis advances the foundational understanding and applications of MAB algorithms in the context of treatment personalization, focusing on improving sample efficiency by leveraging latent structure revealed from historical data, and accommodating practical treatment switching constraints. Key contributions include: (i) latent bandit algorithms for fixed-confidence pure exploration, providing new insights into exploration dynamics; (ii) the Identifiable Latent Bandit framework, which learns reward models from observational data under identifiability assumptions; and (iii) Latent Preference Bandits, which relax structural requirements by modeling preference orderings instead of full reward vectors. The work addresses the challenge of switching constraints through batched exploration approaches. Furthermore, the Alzheimer's Disease Causal estimation Benchmark (ADCB), a semi-synthetic simulator integrating real-world Alzheimer's data with domain expertise is designed and employed as a causally sound evaluation platform for bandit algorithms in personalized medicine. Together, these contributions connect theoretical MAB developments with clinically motivated constraints, offering methodologies for more efficient and practical treatment personalization.

Keywords

Treatment personalization, multi-armed bandits, fixed-confidence pure exploration, latent bandits, structural priors, policy learning with historical data, exploration with switching constraints, healthcare bandit simulators

List of Publications

Appended publications

This thesis is based on the following manuscripts:

I. **Newton Mwai**, Emil Carlsson, Fredrik D. Johansson, *Fast Treatment Personalization with Latent Bandits in Fixed-Confidence Pure Exploration Transactions on Machine Learning Research Journal. April 2023.*

II. Ahmet Zahid Balcıoğlu, **Newton Mwai**, Emil Carlsson, Fredrik D. Johansson, *Identifiable Latent Bandits: Leveraging observational data for personalized decision-making*
In submission, under review. 2025. arXiv preprint: arXiv:2407.16239.

III. **Newton Mwai**, Emil Carlsson, Fredrik D. Johansson, *Latent Preference Bandits*
In submission, under review. 2025. arXiv preprint: arXiv:2508.05367.

IV. **Newton Mwai**, Milad Malekipirbazari, Fredrik D. Johansson, *Understanding exploration in bandits with switching constraints: A batched approach in fixed-confidence pure exploration*
European Workshop on Reinforcement Learning. September 2025.
Also presented at ICML 2024 workshop: Aligning reinforcement learning experimentalists and theorists. 2024.

V. **Newton Mwai**, and Fredrik D. Johansson, *ADCB: An Alzheimer's disease simulator for benchmarking observational estimators of causal effects.*
Conference on Health, Inference, and Learning, pp.103-118, PMLR. 2022.
Also presented at Machine Learning for Health (ML4H). 2021.

Other publications

The following manuscripts were also published during my PhD studies. However, they are not appended to this thesis, due to contents not being related to the thesis.

VI. Lena Stempfle, Anton Matsson, **Newton Mwai**, and Fredrik D. Johansson, *Prediction Models That Learn to Avoid Missing Values. International Conference on Machine Learning (ICML). 2025.*.

VII. Christoffer Ivarsson Orrelid, Oscar Rosberg, Sophia Weiner, Fredrik D. Johansson, Johan Gobom, Henrik Zetterberg, **Newton Mwai**, Lena Stempfle, *Applying machine learning to high-dimensional proteomics datasets for the identification of Alzheimer's disease biomarkers. Fluids and Barriers of the CNS, 22(1), 23. 2025.*

Contribution Summary

I. Co-designed the study, performed most of the empirical work, contributed to the data analysis, and co-wrote the manuscript.

II. Co-designed the study, performed empirical work, contributed to theoretical work, contributed to data analysis, and wrote parts of the manuscript.

III. Co-designed the study, performed most of the empirical work, contributed to theoretical work, contributed to the data analysis, and co-wrote the manuscript.

IV. Co-designed the study, performed most of the empirical work, contributed to the data analysis, contributed to most of the theoretical work, and co-wrote the manuscript.

V. Co-designed the study, performed most of the empirical work, contributed to the data analysis, and co-wrote the manuscript.

Acknowledgements

I am deeply indebted to my PhD advisor, Fredrik D. Johansson, whose unwavering guidance, insightful advice, and genuine intellectual generosity have profoundly shaped my research journey. I have greatly valued his inspiring way of challenging ideas, encouraging deep and creative thinking, and engaging enthusiastically in exploring the finer details of a problem until true clarity is achieved. I am also grateful to my co-advisor Morteza Chehreghani and my examiner Devdatt Dubhashi for their invaluable insights, support and encouragement throughout my PhD studies.

I am fortunate to have collaborated with inspiring co-authors during my PhD: Emil, Ahmet, Lena, Anton, Milad, Christoffer, Oscar, Sophia, Johan, and Henrik. Working with you has been a privilege. Special thanks to Anton, Lena, and Adam, together with whom I became one of the founding PhD students of the Healthy AI Lab, for building a wonderful research environment and a community of camaraderie and great memories. To all my other colleagues at DSAI who I haven't mentioned by name: I consider myself truly lucky to have worked with you, and I deeply value the stimulating exchanges, collaborations, chats, lunches, and countless fikas we've shared along the way.

To my friends and family: thank you for the laughter, support, and grounding you have given me. Special mention to Anna and Anders for your warmth and kindness. To all my family back in Kenya, I cannot fully express the depth of my gratitude. To my mum Mumbi, my sister Njoki, and my brother Kim, thank you for always being present, no matter the distance. Finally, deepest thanks to my dear Katarina, whose light, patience, and love have been a source of strength, even during the most demanding moments of my PhD.

This work was supported in part by WASP (Wallenberg AI, Autonomous Systems and Software Program) funded by the Knut and Alice Wallenberg foundation. Thank you for the invaluable support and all the wonderful experiences through the WASP graduate school. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Thank you.

Contents

Abstract iii

List of Publications v

Acknowledgements vii

I Summary 1

1 Introduction 3

2 Multi-armed Bandits 7
2.1 Regret Minimization in Bandits 8
2.1.1 Bandit Algorithms for Regret Minimization 8
2.1.1.1 Upper Confidence Bound (UCB) 9
2.1.1.2 Thompson Sampling 9
2.2 Pure Exploration in Bandits 11
2.2.1 Fixed-confidence Pure Exploration 11
2.3 Contextual Bandits for Personalization 12

3 Leveraging Latent Bandits to improve Sample Efficiency using Historical Data 15
3.1 Latent Bandits 16
3.2 Challenge 1: Latent bandits in fixed-confidence pure exploration 17
3.3 Challenge 2: Learning identifiable reward models for latent bandits 20
3.4 Challenge 3: Generalizing latent bandits to use looser latent structures of latent preference orderings 22

4 Understanding Bandits with Switching Constraints in Fixed-confidence Pure Exploration 27

5 Semi-synthetic Causal Benchmark for Evaluating Treatment Personalization algorithms 33

6 Conclusion 37

Bibliography 39

II Appended Papers **47**

Paper I - Fast Treatment Personalization with Latent Bandits in Fixed-Confidence Pure Exploration

Paper II - Identifiable Latent Bandits: Leveraging observational data for personalized decision-making

Paper III - Latent Preference Bandits

Paper IV - Understanding exploration in bandits with switching constraints: A batched approach in fixed-confidence pure exploration

Paper V - ADCB: An Alzheimer's disease simulator for benchmarking observational estimators of causal effects.

Part I

Summary

Chapter 1

Introduction

Recent advances in machine learning have fueled growing interest its potential to revolutionize personalized medicine, which seeks to tailor treatments to the unique needs of individual patients. In chronic conditions such as Alzheimer's Disease (AD) (Blennow, de Leon & Zetterberg, 2006; Alzheimer's Association, 2024), Rheumatoid arthritis (RA) (Aletaha & Smolen, 2018; Fraenkel et al., 2021), or Psoriasis (Raharja, Mahil & Barker, 2021; Kim, Jerome & Yeung, 2017) care is an ongoing process that often spans many years. For example, RA presents dozens of therapeutic options following diagnosis (Singh et al., 2016; Murphy, Collins & Rush, 2007), with efficacy varying unpredictably across patients, requiring sequential trials to identify the most effective match. This highlights the critical need for effective treatment personalization strategies. To illustrate the interactive and sequential nature of the treatment personalization process, consider the example of AD treatment outlined in Example 1.

Example 1: Personalized treatment in Alzheimer's Disease (AD): A patient with AD visits a medical clinic seeking treatment. A physician takes diagnostic tests to evaluate patient characteristics related to the disease progression. It is known that AD cannot be cured, but there exist treatments that manage the symptomatic effects manifested with cognitive function (Farlow, Miller & Pejovic, 2008; Livingston et al., 2017; Grossberg et al., 2019). After taking tests, the physician recommends treatments to improve cognitive function, from an available set of treatments, given the patient characteristics. This process is *interactive* because the utility of the recommended treatment (improvement of cognitive function) can only be observed if the patient takes it. Furthermore, since treatments only alleviate symptoms short-term (here cognitive function), this process will have to be repeated *sequentially* at multiple visits, with the physician adjusting treatment, aiming for the patient to achieve the best cognitive function in the long term.

A sequential and interactive framework for exploring alternative options is

multi-armed bandits (MABs) originally motivated by medical applications in drug testing (Thompson, 1933) and studied historically in experiment design (Chernoff, 1959, 1967; Gittens & Dempster, 1979; Lai & Robbins, 1985). MABs have a recent history of personalization applications, with both academic and commercial success (Li et al., 2010; Chapelle & Li, 2011; Bouneffouf, Rish & Aggarwal, 2020a; Yancey & Settles, 2020; O'Brien et al., 2022). In MABs, an agent aims to select actions (e.g. chronic AD treatments) sequentially according to a policy (e.g., a treatment strategy) over multiple rounds, with the goal of maximizing cumulative rewards (e.g. patient cognitive function) from the actions selected. Contextual bandits (Li et al., 2010; Chu et al., 2011; Agrawal & Goyal, 2013; Zhou, 2015; Abbasi-Yadkori, Pál & Szepesvári, 2011) are a well suited bandit formulation for personalization, that extends MABs by incorporating instance-specific covariates (contexts) to tailor decisions, enabling generalization across similar instances and decision points. However, key challenges persist in applying MABs to personalized medicine.

A primary issue is *sample efficiency*: traditional algorithms require extensive interaction steps with an instance to identify optimal actions. This level of per-patient exploration is impractical in healthcare, where the number of treatment opportunities for an individual is limited and the stakes of suboptimal treatment can be high (Dulac-Arnold, Mankowitz & Hester, 2019; Riachi et al., 2021). Efforts to improve sample efficiency in multi-armed bandit algorithms include leveraging historical data for warm-starting learning (Zhang et al., 2019; Oetomo et al., 2023) or learning structures like clusters (Bui, Johari & Mannor, 2012; Bouneffouf et al., 2019; Maillard & Mannor, 2014; Hong et al., 2020a). Latent bandits (Maillard & Mannor, 2014; Hong et al., 2020a; Zhou & Brunskill, 2016; Hong et al., 2020b; Galozy & Nowaczyk, 2023), which assume that each bandit instance belong to unobserved discrete cluster types, have shown theoretical and empirical sample efficiency gains. However, gaps remain in understanding how they reduce exploration time, particularly in pure exploration settings where the goal is to identify the optimal action with high confidence in minimal trials (Garivier & Kaufmann, 2016; Kaufmann, 2020), practical learning from historical data (Agrawal et al., 2023), and adaptability to looser information structures.

Another challenge with MABs in personalized medicine is incorporating practical constraints, such as minimizing treatment switches to reduce patient burden or adhering to clinical guidelines. In MAB literature, switching has been addressed extensively in regret minimization. For example Arora, Dekel and Tewari (2012) studied switching in regret minimization against adaptive adversaries and showed how mini-batching can control switches while keeping regret low. Similar approaches appear in Dekel et al. (2014), Rouyer, Seldin and Cesa-Bianchi (2021), Amir et al. (2022) and Li et al. (2023). However, controlling for switching is under-explored in the pure exploration setting, where the goal is not to maximize cumulative reward but to identify the optimal arm efficiently.

Moreover, evaluating MAB algorithms in personalized medicine is challenging. Live testing of MAB algorithms in clinical settings is rarely feasible due to ethical risks, patient safety concerns, regulatory constraints, and more.

Simulators provide a safe and controlled environment to explore algorithm performance without endangering patients or violating regulatory protocols. For such simulators to be useful, preserving causal relationships between clinical variables is critical. Without causal fidelity, algorithm evaluations may be misleading, particularly in sequential treatment settings where decision outcomes depend on dynamic, interrelated factors. Yet, realistic benchmarks that combine causal realism with healthcare complexity remain scarce. While simulators like IHDP (Hill, 2011) and ACIC (Dorie et al., 2019) have been valuable for causal effect estimation, they rely on simplified, static mathematical response surfaces and are not designed for sequential decision-making as in bandit settings. More data-driven approaches (Chan et al., 2021; Neal, Huang & Raghupathi, 2020; Kuo et al., 2022) increase realism but often overlook underlying causal mechanisms. This underscores the need for hybrid benchmarks that integrate real clinical data with domain-expert causal knowledge (Hernán, 2019), enabling both realism and validity in evaluating bandit algorithms in personalized medicine.

This thesis addresses the mentioned gaps by studying and developing novel MAB-based strategies motivated by applications in personalized medicine, particularly in chronic diseases. MABs are reviewed in Chapter 2, and later chapters introduce results and approaches in:

i. Formulating latent bandits in fixed-confidence pure exploration, revealing how latent structures reduce sample complexity by shrinking alternative parameter sets, and proposing asymptotically optimal algorithms (Chapter 3).

ii. Introducing the Identifiable Latent Bandit framework, proving identifiable latent variable models from historical data under causal assumptions enable optimal online decision-making (Chapter 3).

iii. Proposing Latent Preference Bandits (LPB), which use preference orderings for looser latent structures, and demonstrating characteristics and utility of LPB empirically (Chapter 3).

iv. Reformulating pure exploration with switching constraints using batched plays, developing optimal algorithms that minimize batches while limiting switches and presenting optimality guarantees (Chapter 4).

v. Designing the ADCB simulator, a semi-synthetic benchmark for Alzheimer's disease that integrates real Alzheimer's disease (AD) data with causal domain knowledge and treatments from AD literature, for robust evaluation of MABs in the AD setting (Chapter 5).

Chapter 2

Multi-armed Bandits

Multi-armed bandits (*MABs* or *bandits*) were introduced as a framework for exploring alternative options, originally motivated by reducing suffering in drug testing (Thompson, 1933) and subsequently studied in various settings in experiment design and stochastic adaptive allocation (Chernoff, 1959, 1967; Gittens & Dempster, 1979; Lai & Robbins, 1985). The name is inspired by a casino game where a gambler aims to find the best slot machine of "one-armed bandit" casino machines through repeated trials. Bandits are a precise and efficient toolbox for a large class of problems in the standard model of reinforcement learning (Bouneffouf & Féraud, 2024) and they've seen a renewed research interest since the demonstration of their application in personalized news recommendation (Li et al., 2010) with successful study and application in varied domains from healthcare to finance, and more (Bouneffouf, Rish & Aggarwal, 2020b).

A multi-armed bandit problem is defined as follows (Lattimore & Szepesvári, 2020): An *agent* (e.g., a treatment personalization strategy) and an *environment* (e.g., a patient) interact in sequence over T rounds. For each round t, the agent takes an *action* $A_t \in \mathcal{A} = \{1, \ldots, K\}$ (e.g., a treatment) and receives a *reward* $R_t \in \mathbb{R}$ (e.g., a treatment outcome). Formally, a stochastic bandit setting comprises a set of K actions, $\mathcal{A} = \{1, \ldots, K\}$, and the environment specifies K reward probability distributions $\nu_1, \ldots, \nu_K$ with their respective parameters (e.g. for Gaussian distributions with the same variance σ^2: the means $\mu \in \mathbb{R}^K$). An agent seeks to learn the reward distributions of these arms by interacting with the environment, as illustrated in Algorithm 1.

Algorithm 1 Multi-Armed Bandit Problem

1: **for** each round $t = 1, \ldots, T$ **do**
2: $a_t \leftarrow$ Agent chooses an arm $a_t \in \mathcal{A}$ using an exploration-exploitation strategy (algorithm) $\pi(\hat{\mu})$ based on current parameter estimates $\hat{\mu}$
3: $r(a_t) \leftarrow$ A new independent, stochastic reward r_t is realized, drawn from ν_{a_t}
4: Update estimated parameters $\hat{\mu}$
5: **end for**

The key challenge in the MAB problem is the *exploration–exploitation dilemma*: deciding when to explore (select actions whose rewards are not yet well-estimated) versus when to exploit (select the action currently believed to yield the highest reward). The difficulty arises because the true reward distributions $\nu_a,\ a \in \mathcal{A}$ are unknown to the agent (here, "agent" and "algorithm" are used interchangeably) at the start of the interaction. The agent must estimate the parameters $\hat{\mu}$ from observed rewards, and these estimates remain uncertain even with a large number of rounds T (Elena, Milos & Eugene, 2021).

In estimating these distributions, there are various objectives that are typically formulated, that are next introduced.

2.1 Regret Minimization in Bandits

The typical goal in MAB studies is to design algorithms π that select actions $a \in \mathcal{A} = \{1, ..., K\}$ to maximize the cumulative reward over rounds. This is a setting referred to as *regret minimization*. The regret minimization goal is defined with respect to the unknown expectations of rewards $\mu_a \coloneqq \mathbb{E}[R_a]$, with the optimal action $a^* = \arg\max_a \mu_a$ and optimal reward $\mu^* = \mu_{a^*}$. The aim is to select actions a_t according to an algorithm π on times steps $t = 1, ..., T$ until a horizon T to accumulate as little regret $\mathrm{Reg}(T)$ as possible,

$$\underset{\pi}{\text{minimize}}\ \mathrm{Reg}(T) \quad \text{with} \quad \mathrm{Reg}(T) \coloneqq \sum_{t=1}^{T} \mathbb{E}_\pi[\mu^* - R_{A_t}]\ . \tag{2.1}$$

To evaluate the performance of a MAB algorithm, *regret bounds* are typically used. A *regret lower bound* characterizes the inherent difficulty of a bandit problem by stating the minimum regret that *any* algorithm must incur in the worst case, given a specified class of environments and a time horizon (Lattimore & Szepesvári, 2020; Salomon, Audibert & Alaoui, 2011). These bounds are algorithm-agnostic, and they hold for all possible algorithms within the model class, such as all *consistent* policies in the stochastic bandit setting. In contrast, an *upper bound* applies to a specific algorithm (or family of algorithms) and shows that its regret does not exceed a certain level. When an algorithm's upper bound matches the lower bound (up to constant or logarithmic factors), it is considered minimax optimal.

In stochastic regret minimization, a fundamental result by Auer et al. (1995) establishes a minimax lower bound of order $\Omega(\sqrt{KT})$ on the expected regret. That is, for any bandit algorithm, there exists at least one problem instance in which the expected regret satisfies $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$. This bound highlights the inherent difficulty of the problem and serves as a benchmark for evaluating algorithms, motivating the introduction of several bandit strategies whose performance matches this rate up to logarithmic factors.

2.1.1 Bandit Algorithms for Regret Minimization

A naïve regret minimization strategy is to always select the arm that seems based on the current knowledge. This would be, for instance, the arm with the highest estimated reward after the first K trials. The drawback with such

a purely greedy strategy is the lack of exploration. While it may perform well in the short-term, it may not necessarily find the optimal solution in the long term (Bouneffouf, 2023). To overcome this, bandit strategies explicitly incorporate exploration, and aim to balance the trade-off between sampling arms with uncertain rewards (exploration) and selecting the empirically best arm so far (exploitation). Two of the most widely used approaches in stochastic bandits that embody this principle are the Upper Confidence Bound (UCB) and Thompson Sampling (TS) approaches, which are introduced next.

2.1.1.1 Upper Confidence Bound (UCB)

The idea behind the Upper Confidence Bound (UCB) class of algorithms (Lai, Robbins et al., 1985; Agrawal, 1995; Auer, Cesa-Bianchi & Fischer, 2002; Auer, 2002) is centered around the *principle of optimism in the face of uncertainty*, which states that in environments with uncertainty, it is beneficial for the agent to assume that the environment is as favorable as plausibly possible, and to act accordingly. UCB algorithms compute an optimistic estimate of each arm's reward by combining the empirical mean with a confidence bonus, and then select the arm with the highest such optimistic estimate. An example is the `UCB1` strategy (Auer, Cesa-Bianchi & Fischer, 2002), which at each time step t chooses the arm a that maximizes the following value:

$$\mathrm{UCB}_a(t) = \underbrace{\hat{\mu}_a(t-1)}_{\text{estimate}} + \underbrace{\sqrt{\frac{2\log(t)}{N_a(t-1)}}}_{\text{confidence bound}} \tag{2.2}$$

Here, $\hat{\mu}_a(t-1)$ represents the estimated mean reward of arm a based on the outcomes of all previous pulls up to time $t-1$. This term encourages the selection of arms that have historically provided high rewards, thus promoting exploitation. The second term, $\sqrt{\frac{2\log(t)}{N_a(t-1)}}$, serves as a confidence bound that quantifies the uncertainty in the estimate of $\hat{\mu}_a(t-1)$, which is ensured to be an upper bound for the unknown means with high probability. $N_a(t-1)$ denotes the number of times arm a has been selected up to time $t-1$, and $\log(t)$ is the natural logarithm of the total number of pulls across all arms up to time t.

The confidence bound is larger for arms that have been selected fewer times (i.e., when $N_a(t-1)$ is small), thereby increasing the UCB value for those arms and making them more likely to be chosen. This mechanism ensures that the algorithm explores arms with greater uncertainty, as their true mean rewards might be higher than currently estimated. By selecting the arm with the highest UCB value at each time step, the algorithm effectively balances the desire to exploit arms with high estimated rewards and the need to explore arms with potential for higher rewards due to uncertainty in their estimates.

2.1.1.2 Thompson Sampling

Thompson Sampling (Thompson, 1933), also known as *posterior sampling*, is a classical Bayesian method for solving multi-armed bandit (MAB) problems.

Algorithm 2 Upper Confidence Bound (`UCB1`) Algorithm

Require: the number of arms K
1: **for** $t = 1$ to K **do**
2: Play arm t
3: **end for**
4: **for** $t = K+1, K+2, \dots T$ **do**
5: Select arm $a_t = \arg\max_{a \in [K]} \left(\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(t)}{N_a(t-1)}} \right)$
6: **end for**

For each arm a, the reward $r_a(t)$ at time t is modeled by a distribution $P(r_t|\hat{\lambda}_a)$, where $\hat{\lambda}_a$ represents the parameters of the reward distribution for arm a. A prior distribution $P(\hat{\lambda}_a)$ is specified for each arm's parameters. After observing a reward r_t from playing an arm, the posterior $P(\hat{\lambda}_a|r_t)$ is updated using Bayes' rule: $P(\hat{\lambda}_a|r_t) \propto P(r_t|\hat{\lambda}_a)P(\hat{\lambda}_a)$, based on all rewards observed for that arm up to that point. At each time step t, the algorithm samples parameters $\tilde{\lambda}_a$ from the current posterior distribution of each of the K arms, computes the expected reward $\mathbb{E}[r_a|\tilde{\lambda}_a]$ for each arm given these samples, and selects the arm with the highest expected reward. The following algorithm illustrates this process: Thompson Sampling (TS) has been shown to be a

Algorithm 3 Thompson Sampling Algorithm

Require: the number of arms K, prior distributions $P(\hat{\lambda}_a)$ for each arm $a = 1, \dots, K$
1: Initialize: For each arm a, set $P_{\text{post},a} \leftarrow P(\hat{\lambda}_a)$
2: **for** $t = 1, 2, \dots T$ **do**
3: For each arm a, sample $\tilde{\lambda}_a \sim P_{\text{post},a}$
4: Select arm $a_t = \arg\max_{a \in [K]} \mathbb{E}[r_a|\tilde{\lambda}_a]$
5: Play arm a_t
6: Observe reward r_t
7: Update $P_{\text{post},a_t} \leftarrow \text{Update}(P_{\text{post},a_t}, r_t)$
8: **end for**

competitive and often high-performing approach to stochastic MAB problems, frequently matching or outperforming algorithms such as UCB across diverse application domains (Chapelle & Li, 2011; Graepel et al., 2010). Theoretically, analyses (Agrawal & Goyal, 2012; Kaufmann, Korda & Munos, 2012; Russo & Van Roy, 2014) have established that TS achieves asymptotically optimal regret in the stochastic setting. Moreover, variants of both UCB and TS enjoy worst-case (minimax) guarantees of $\mathbb{E}[R(T)] \leq O(\sqrt{KT \log T})$, matching the minimax lower bound up to logarithmic factors (Lattimore & Szepesvári, 2020). While these results do not prove that TS is uniformly superior to UCB in all settings, they demonstrate that both algorithm families combine strong empirical performance with robust theoretical guarantees in both problem-dependent and worst-case regimes.

2.2 Pure Exploration in Bandits

While many bandit problems focus on maximizing cumulative reward by balancing exploration and exploitation, some formulations remove the incentive to exploit altogether. These are known as *pure exploration* (PE) problems, in which the agent's objective is to gather information about the environment as efficiently as possible, regardless of the rewards (Kaufmann, 2020). In the PE setting, two main problem types are commonly studied: (i) the *fixed-confidence setting*, where the goal is to identify the best arm using the fewest possible rounds while achieving a pre-specified probability of success; and (ii) the *fixed-budget setting*, where the goal is to identify the best arm with the highest possible probability of success given a fixed number of rounds. This thesis focuses on the fixed-confidence pure exploration setting, outlined next.

2.2.1 Fixed-confidence Pure Exploration

A fixed-confidence pure-exploration strategy ϕ comprises a sampling rule for exploring actions A_t at each step t, a stopping rule to decide the time τ at which the exploration is over, and a recommendation rule which returns the best action $\hat{a}_\tau$ at the stopping time τ (Garivier & Kaufmann, 2016; Kaufmann, 2020; Shang et al., 2020). The goal is usually to design a strategy ϕ to minimize the expected stopping time $\mathbb{E}[\tau]$ with a pre-specified confidence parameter δ:

$$\begin{aligned} &\underset{\phi}{\text{minimize}} && \mathbb{E}_\phi[\tau] \\ &\text{subject to} && P(\mu_{\hat{a}_\tau} < \mu^*) \leq \delta, \end{aligned} \tag{2.3}$$

For the fixed-confidence pure exploration problem, Garivier and Kaufmann, 2016 presented a general lower bound for the expected stopping time $\mathbb{E}[\tau]$ of any δ-PAC multi-armed bandit algorithm, i.e., one that returns the best arm with probability at least $1-\delta$, for some $\delta > 0$,

$$\mathbb{E}[\tau] \geq T^*(\mu)\ \mathrm{kl}(\delta, 1-\delta)\ . \tag{2.4}$$

$$\text{where}\quad T^*(\mu)^{-1} := \sup_{w \in \Sigma^K} \inf_{\lambda \in \mathrm{Alt}(\mu)} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \right).$$

Here, $d(.)$ is the KL-divergence, and $\Sigma^K := \{w \in \mathbb{R}_+^K : \sum_{a=1}^K w_a = 1\}$ is the simplex of possible arm playing proportions. This lower bound is derived by considering the optimal allocation of arm pulls w^* to minimize the worst-case stopping time specific to the instance μ while ensuring that the probability of incorrectly identifying the best arm does not exceed a pre-specified confidence level δ. The term $T^*(\mu)$ is a "characteristic time" for the problem, that depends on the parameters of the arms. It represents the inverse of the exploration time associated by the best-case (supremum) playing proportions w and the worst-case (infimum) alternative bandit model λ (that differs from μ in its optimal arm), $\mathrm{Alt}(\mu) = \{\lambda \in \mathbb{R}^K : \arg\max_a \lambda_a \neq \arg\max_a \mu_a\}$.

$T^*(\mu)^{-1}$ is the maximum achievable *information acquisition rate*, obtained by choosing the arm allocation w that maximizes the smallest KL-divergence

to any alternative model λ. The characteristic time $T^*(\mu)$ is its inverse: it expresses the minimum number of samples (up to the $\log(1/\delta)$ factor; using $\mathrm{kl}(\delta, 1-\delta) \sim \log(1/\delta)$ in Eq. 2.4) required to identify the optimal arm in instance μ with high confidence. The higher the information acquisition rate $T^*(\mu)^{-1}$, the smaller the characteristic time $T^*(\mu)$ and thus the faster the problem can be solved. From this definition, Garivier and Kaufmann (2016) derive an *asymptotic* lower bound for any δ-PAC algorithm as $\delta \to 0$, using $\mathrm{kl}(\delta, 1-\delta) \sim \log(1/\delta)$:

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

Several strategies including those based on arm elimination, adaptivity, racing and upper-confidence bounds have been proposed for the fixed-confidence pure exploration setting by Garivier and Kaufmann (2016), Kalyanakrishnan et al. (2012), Gabillon, Ghavamzadeh and Lazaric (2012), Jamieson and Nowak (2014), Jun et al. (2016) and Jedra and Proutiere (2020) among others. Of particular interest in this thesis is the adaptive "Track-and-Stop" class of algorithms which originates from the analysis by Garivier and Kaufmann (2016). These algorithms are designed to *track* the optimal arm playing proportions $w^*(\hat{\mu})$ of the lower bound in (2.4),

$$w^*(\hat{\mu}) := \arg\max_{w \in \Sigma^K} \inf_{\lambda \in \mathrm{Alt}(\hat{\mu})} \left(\sum_{a=1}^{K} w_a d(\hat{\mu}_a, \lambda_a) \right) . \tag{2.5}$$

based on an estimate $\hat{\mu}$ of the arm parameters, continuously updated as more data is collected. A track-and-stop algorithm plays arms following a *tracking rule* aiming for an overall arm proportion as close to the optimal proportions as possible, combined with a *stopping rule* for terminating exploration. The stopping rule is a statistical test of whether the past observations indicate, with a risk of at most δ, that one arm has a higher average reward than the others.

2.3 Contextual Bandits for Personalization

Contextual bandits (Li et al., 2010; Chu et al., 2011; Agrawal & Goyal, 2013; Zhou, 2015) extend the classic multi-armed bandit (MAB) framework to address decision-making under uncertainty across diverse scenarios, enabling algorithms that effectively balance exploration *and* generalization. They were originally introduced for internet news personalization by (Li et al., 2010), and they they have seen success in numerous personalization applications since (Chapelle & Li, 2011; Bouneffouf, Rish & Aggarwal, 2020a; Yancey & Settles, 2020; O'Brien et al., 2022).

Unlike traditional MABs, which focus solely on identifying the best action averaged over all situations, contextual bandits determine the optimal action conditional on side information available at each decision point. This is achieved by augmenting the MAB problem with a context variable $X_t \in \mathcal{X}$ (e.g., a vector in $\mathbb{R}^d$) that is observed at the start of each round t. For example, in a

treatment personalization application, the agent may observe patient-specific covariates, such as lab measurements or demographic data, at the beginning of each clinical visit. These per-round contexts allow the agent to select actions tailored to the current situation and to generalize knowledge across similar contexts, thereby improving long-term performance.

A central challenge in contextual bandits is designing reward models that effectively capture the underlying distribution of rewards given contexts and actions. To address this, various assumptions have been proposed (Zhou, 2015; Lattimore & Szepesvári, 2020), such as linearity or Lipschitz continuity of the reward function, enabling the development of contextual bandit strategies for regret minimization.

For instance, in the stochastic linear contextual bandit setting (Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori, Pál & Szepesvári, 2011; Auer, 2002; Langford & Zhang, 2007), the expected reward for an action $a \in \mathcal{A} = \{1, \ldots, K\}$ given a context x_t is modeled as a linear function $f(x_t, a) = \langle \theta^*, \phi(x_t, a) \rangle$, where θ^* is an unknown parameter vector, and $\phi(x_t, a)$ is a known feature mapping. The agent receives a reward $r_t = f(x_t, a_t) + \epsilon_t$, with ϵ_t being independent Gaussian noise with mean 0 and variance 1. The goal is to maximize the expected cumulative reward over T rounds, or equivalently, minimize the pseudo-regret $R_T = \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{A}} f(x_t, a) - \sum_{t=1}^{T} f(x_t, a_t)\right]$. Within this framework, the LinUCB algorithm (Abbasi-Yadkori, Pál & Szepesvári, 2011) achieves an expected regret upper bound $\mathbb{E}[R(T)] \leq Cd\sqrt{T}\log(TL)$, where $C > 0$ is a constant, d is the feature dimension and L is a Lipschitz constant.

The contextual bandit problem proceeds as outlined below in Algorithm 4.

Algorithm 4 Contextual Bandit Problem

1: **for** each round $t = 1, 2, \ldots, T$ **do**
2: Observe context x_t
3: Select action $a_t \in \mathcal{A}$ using an exploration–exploitation strategy based on current parameter estimate $\hat{\theta}$
4: Receive reward r_t drawn from the conditional distribution $P(\cdot \mid x_t, a_t)$
5: Update $\hat{\theta}$ using $\{(x_s, a_s, r_s)\}_{s=1}^{t}$
6: **end for**

Contextual bandits are well-suited for healthcare applications, such as personalized treatment selection, because they can leverage contextual information to adapt decisions to patient-specific characteristics. In healthcare, where patient heterogeneity and limited intervention opportunities make efficient learning critical, this ability to generalize across similar contexts is particularly valuable. However, personalization is only as complete as the information captured in the context: in classical contextual bandits, each instance (patient) is treated independently, and model parameters are typically learned from scratch for each new context. This leads to long exploration periods, which is undesirable in realistic clinical settings, motivating the work in this thesis on sample-efficient treatment personalization algorithms.

Chapter 3

Leveraging Latent Bandits to improve Sample Efficiency using Historical Data

A pragmatic solution to minimize the sample complexity in personalized decision-making is to leverage (offline) observational logs of previous decisions and outcomes for to speed up the online decision process. Logs of decision processes collected over past periods are plentiful in many sequential decision making environments. In healthcare, they are typically collected as electronic health records (EHRs) and abound in many healthcare systems, often covering past records of treatment procedures of multiple patients over long periods (Ambinder, 2005).

There is extensive literature on learning and evaluating policies from logged bandit feedback, often referred to as *off-policy learning* (when the goal is to learn a new policy) or *off-policy evaluation (OPE)* (when the goal is to estimate the value of a given policy) (Strehl et al., 2010; Dudík, Langford & Li, 2011; Swaminathan & Joachims, 2015a, 2015b). In this setting, agents must operate entirely on an offline dataset collected under a potentially unknown logging policy. A key challenge in policy evaluation from logged data is that the logging policy may be non-uniformly stochastic, which can introduce bias in action selection and high variance in value estimates when some action propensities are small (Joachims et al., 2021). Common approaches for addressing these challenges include the *Inverse Propensity Score (IPS)* method (Horvitz & Thompson, 1952; Swaminathan et al., 2017), *Direct Methods* (Beygelzimer & Langford, 2009), and *Doubly Robust* estimators (Dudík, Langford & Li, 2011; Robins & Rotnitzky, 1995). However, these methods generally require strong overlap assumptions on the logging policy's action probabilities, which limits their applicability in practice (Yin & Wang, 2021). Due to these challenges, this thesis focuses on blending online and offline learning by using historical

data to reveal latent structure, rather than learning policies purely from offline data.

In MABs, the main approaches for blending online and offline learning to shorten exploration are either: i) Warm-starting model parameters for online learning, with historical data in an offline phase (Zhang et al., 2019; Oetomo et al., 2023, 2024), or ii) Leveraging historical data to reveal structure about the data through clustering (Bui, Johari & Mannor, 2012; Bouneffouf et al., 2019; Maillard & Mannor, 2014; Hong et al., 2020a; Huch et al., 2024), matrix decomposition (Sen et al., 2017), or spectral methods (Kocák et al., 2020).

Latent bandits (Maillard & Mannor, 2014; Hong et al., 2020a; Zhou & Brunskill, 2016; Hong et al., 2020b), studied in this thesis, assume that each bandit instance belong to unobserved discrete types, and they have proved theoretically and empirically more sample efficient than unstructured bandits. However, they come with unexplored challenges that have been a focus.

3.1 Latent Bandits

Latent bandits (Maillard & Mannor, 2014; Hong et al., 2020a; Zhou & Brunskill, 2016; Hong et al., 2020b) are an extension of contextual bandits where the reward R_t at time t depends on context X_t, action A_t, and a *latent state* $s \in \mathcal{S}$, where s is fixed but unknown for new instances at the start of interaction. The reward is sampled from a *conditional reward distribution*, $P(\cdot \mid A, X, s, \theta)$, which is parameterized by a vector $\theta \in \Theta$, where Θ is the space of feasible reward models. The model parameters θ are typically assumed to be available to the learner in advance in previous works. However, these works do not address the problem of recovering such parameters from data. In this thesis, θ is either assumed to be known, as in the prior literature, or estimated from historical data. When estimated, this is done using an offline interaction log $H_t = (X_1, A_1, R_1, \ldots, X_t, A_t, R_t)$ of contexts, actions, and rewards up to time t, with h_t denoting its observed realization. Under a Gaussian assumption, the mean reward for arm a is $\mu(a, x, s, \theta) = \mathbb{E}_{R \sim P(\cdot|a,x,s,\theta)}[R]$, where s is the latent state and θ is the parameter vector, either known in advance or estimated offline from the logged data. Latent bandits have largely been studied assuming discrete latent states, therefore with $|\mathcal{S}||\mathcal{A}|$ probability distributions $\nu_{s,1}, \ldots, \nu_{s,K}$ with respective parameters $\mu_{s,1}, \ldots, \mu_{s,K}$, which are specified a priori. A latent bandit problem proceeds as follows:

Algorithm 5 Latent Bandit Problem

1: **for** each round $t = 1, 2, \ldots, T$ **do**
2: Algorithm observes a context x_t
3: Algorithm estimates latent state s_t
4: $a_t \in \mathcal{A}$ is chosen using an exploration–exploitation strategy
5: A new independent, stochastic reward r_t is realized, drawn from the distribution $\nu_{s_t,a_t}(x_t)$
6: Updates are made for estimated latent state parameters $\hat{\theta}$ in $p(s \mid h_t, \hat{\theta})$
7: **end for**

Hong et al. (2020a) provide algorithms and propose regret upper bounds $\mathbb{E}[R(T)] \leq O(\sqrt{MT\log(T)})$ which depend on the latent state dimension M. The informativeness of this bound will be brought into question in this thesis in Section 3.4.

3.2 Challenge 1: Latent bandits in fixed-confidence pure exploration

While latent bandits have been shown to be empirically more sample efficient than other traditional bandits, key questions arise relating to: i) How they leverage the latent structure to achieve sample efficiency *during exploration*, and ii) what the fundamental limits of exploration are, if the goal is to obtain an optimal arm. These questions can be answered precisely in the fixed-confidence pure exploration problem formulation, where the goal is to identify the optimal action $a^* = \arg\max_a \mu_{a,x,s}$ with confidence $1-\delta$ in minimal expected time $\mathbb{E}_\phi[\tau]$. However, while regret minimization (RM) in latent bandits is well-studied, latent bandits have not been explored in the fixed-confidence pure-exploration (FC-PE) setting.

In Paper I (Kinyanjui, Carlsson & Johansson, 2023), we formulate the fixed-confidence pure-exploration latent bandit problem, where an agent observes a context x, takes actions A_t, observes rewards $R_t \sim \mathcal{N}(\mu_{a_t,x,s}, \sigma^2)$, and stops at time τ to recommend $\hat{a}_\tau$. The objective is to design a search strategy ϕ to minimize $\mathbb{E}[\tau]$ subject to $P(\mu_{\hat{a}_\tau,x,s} \neq \mu^*_{x,s} \mid X = x, S = s) \leq \delta$. We assume a finite number of latent states $S \in \mathcal{S} = \{1, ..., M\}$, and stationarity in the latent states, as in previous work (Hong et al., 2020a). Given that our scope is only in understanding *exploration with latent structure*, we assume that the conditional reward models are known. In the context of this work, we define a latent variable model (LVM) as $\mathcal{M}_\theta = \{p_\theta(s), p_\theta(x \mid s), p_\theta(r \mid a, x, s)\}$, which we assume is available, implying conditional reward models are available. The LVM informs the posterior $p(s \mid h_t)$ that guides the estimation of the unknown latent state in the latent bandit problem. This is illustrated in Figure 3.1.

By the change of distribution argument (Lai, Robbins et al., 1985) with the alternate set $\text{Alt}_x(s) := \{s' \in \mathcal{S} : \arg\max_a \mathbb{E}[r|s, x, a] \neq \arg\max_a \mathbb{E}[r|s', x, a]\}$, and with the help of the information-processing lemma (Garivier & Kaufmann, 2016; Thomas M. Cover, 2005), we derive a lower bound on the expected stopping time for any δ-PAC algorithm (Proposition 1):

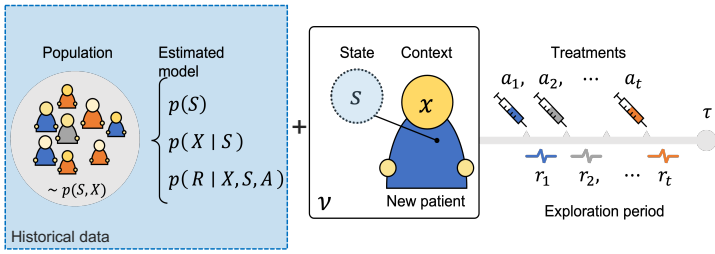


Figure 3.1: Illustration of the pure-exploration latent bandit problem and the example of treatment personalization. A population of patients have been observed in historical data to learn the distribution of latent states $P(S)$, $P(X|S)$ and the conditional reward the distribution $P(R|X,S,A)$. A new patient, represented by the instance $\nu = (x, s)$ is treated with actions a_t, observing rewards r_t until the stopping time τ. The goal is to understand how latent bandits leverage the latent structure to achieve sample efficiency *during exploration*, and what the fundamental limits of exploration are.

Proposition 1 *For any δ-PAC learner ϕ with $\delta \in (0, 1/2)$ and any latent state s and context x, the expected stopping time satisfies*

$$\mathbb{E}_\phi[\tau \mid s, x] \geq \frac{1}{C^*_\delta(s,x)} \mathbf{kl}(\delta||1-\delta)$$

where $1/C^*_\delta(s,x) = \sum_a \gamma^*_{x,a}(s)$ with $\gamma^*_{x,a}(s)$ the minimizers of the following linear program,

$$\underset{\gamma_{x,a} \geq 0}{\text{minimize}} \sum_a \gamma_{x,a} \tag{3.1}$$

$$\text{subject to} \sum_a \gamma_{x,a} \mathbb{KL}^{R,a,x}_{s,s'} + \frac{\rho(x; s, s')}{\mathbf{kl}(\delta||1-\delta)} \geq 1, \ \forall s' \in \text{Alt}_x(s)$$

where $C^*_\delta(s,x) = \sum_a \gamma^*_{x,a}(s)$ is a sample complexity term ("characteristic time"), and $\gamma^*_{x,a}(s)$ are solutions to a linear program (LP) minimizing exploration under KL-divergence constraints connecting the optimal worst-case solution to our exploration objective with hardness of separation of latent states s, s'. A bound for the population (marginal) search time follows as

$$\mathbb{E}_{\phi,X,S}[\tau] \geq \frac{1}{C^*_\delta} \mathbf{kl}(\delta||1-\delta),$$

assuming that $\frac{1}{C^*_\delta} = \mathbb{E}_{X,S}[\sum_a \gamma^*_{x,a}(s)]$ exists, with $\gamma^*_{x,a}$ the minimizers as before.

Because RM and FC-PE objectives are different, algorithms are not directly transferrable between these settings. We therefore propose FC-PE latent bandit algorithms, i) the Latent LP-based Track and Stop (LLPT) Explorer, which tracks optimal arm proportions, $w^*_{x,a}(s) = \gamma^*_{x,a}(s)/(\sum_a \gamma^*_{x,a}(s))$, from the lower bound linear program, and ii) the Divergence Explorer, which selects actions

maximizing expected KL-divergence between latent states. The LLPT Explorer matches this bound asymptotically in the high-confidence limit ($\delta \to 0$), proving optimality (Proposition 2).

Proposition 2 *Let τ be the stopping time of LLPT Explorer ϕ. With s the true state and $C^*(s,x)$ the optimum in (3.1) with the ρ-term removed, there is a constant $\alpha > 0$ such that*

$$\lim_{\delta \to 0} \frac{\mathbb{E}_\phi[\tau \mid s, x]}{\log(1/\delta)} \leq \frac{\alpha}{C^*(s,x)} \,. \tag{3.2}$$

Empirical validation on an Alzheimer's disease simulator (ADCB) (Kinyanjui & Johansson, 2022) shows that both algorithms have a significantly reduced sample complexity compared to baselines oblivious of latent structure like Top-Two Thompson Sampling (TTTS) (Russo, 2016) as seen in Figure 3.2.

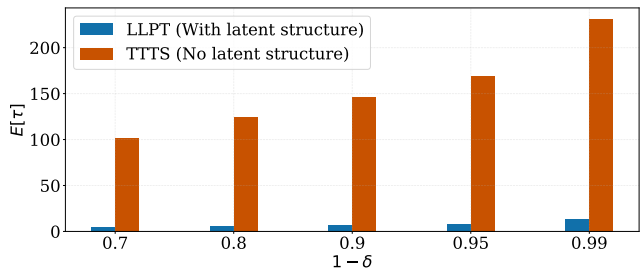


Figure 3.2: Using latent state structural information significantly reduces the expected number of trials $\mathbb{E}[\tau]$ required to identify an optimal treatment with confidence at least $1 - \delta$ in a simulator of Alzheimer's disease progression.

Looking into this challenge reveals fundamental insights on exploration in latent bandits. Our key result is demonstrating that the optimal worst-case solution to the exploration objective relates to hardness of separation of latent states s, s' (i.e similarity of latent states) theoretically and empirically (See Figure 3.3). Another insight explaining sample efficiency with latent structure is that the sample complexity term $C^*(s,x)$ shrinks when we have knowledge of the latent state structure because the set of plausible alternative parameters $\text{Alt}_x(s)$ is smaller compared to the case with no structure in, for example, Garivier and Kaufmann (2016). In latent bandits, $\text{Alt}_x(s)$ comprises a finite set of parameters, whereas the case where parameters are estimated online without latent structure corresponds to an infinite set of alternative parameters. As a result, the worst-case (supremum) over alternative parameter sets shrinks, as do the lower and upper bounds on the stopping time. In RM, despite a different objective, it is reasonable to assume that exploration is similarly characterized, and the insights transferrable.

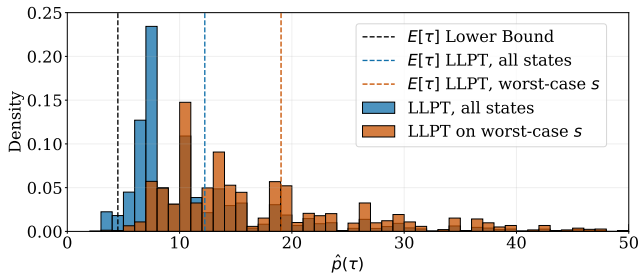


Figure 3.3: Density of stopping times under LLPT showing worst-case latent states revealing that higher stopping times can be attributed to the worst-case latent states, i.e, exploration difficulty depends on the distinguishability of latent states.

3.3 Challenge 2: Learning identifiable reward models for latent bandits

Another challenge relates to the limitation of assuming known conditional reward models. The key components of latent bandit algorithms are a latent variable model (LVM) approximating $p(Z_i \mid H_{i,t}, X_{i,t})$ and a reward model $\mu_a(z)$ for each value of z. The reward model is used to select the next action according to a *selection criterion* based on an inferred value of Z. Here, Z denotes the random variable representing the latent state; $i \in [I]$ indexes previous problem instances, each with a sequence length T_i, and $t \in [T_i]$ indexes rounds within an instance. For example, the **mTS** algorithm (Hong et al., 2020a) samples $\hat{z}_t \sim p(Z_i \mid H_{i,t}, X_{i,t})$ and selects the action $a_t = \arg\max_a \mu_a(\hat{z}_t)$. However, this and related works assume that both state and reward models are known a priori, but give little guidance for how to learn or acquire them. To make real-world application plausible, algorithms must learn the LVM from *observational historical data* $\mathcal{D} = \{(x_{1,t}, a_{1,t}, r_{1,t})_{t=1}^{T_1}, ..., (x_{I,t}, a_{I,t}, r_{I,t})_{t=1}^{T_I}\}$. This presents a new problem: not all LVMs are *identifiable*, as they may fail to recover the true underlying process that generated $\mathcal{D}$ (Hyvarinen & Morioka, 2016). So a question that arises is: How can identifiable LVMs be learned from historical data and can identifiable LVMs be shown to provably yield optimal decision making in latent bandits?

In Paper II (Balcıoglu et al., 2025), we propose the Identifiable Latent Bandit (ILB) framework, which combines offline learning of a latent variable model with online decision-making to minimize regret. The focus is in learning an identifiable LVM, but we also contextualize the problem in a latent bandit (with a *continuous* latent state) regret minimization setting, and investigate identifiability of decision-making in ILB within a causality framework (Pearl, 2009). We learn the LVM offline, by starting with an assumed structural causal model illustrated in Figure 3.4, and identifiably learn the inverse emission function g^{-1} and reward model θ from the observational data $\mathcal{D}$, using contrastive learning with multinomial logistic regression (Hyvarinen & Morioka, 2016), to support inferring the latent state Z_i and the best possible action for a new

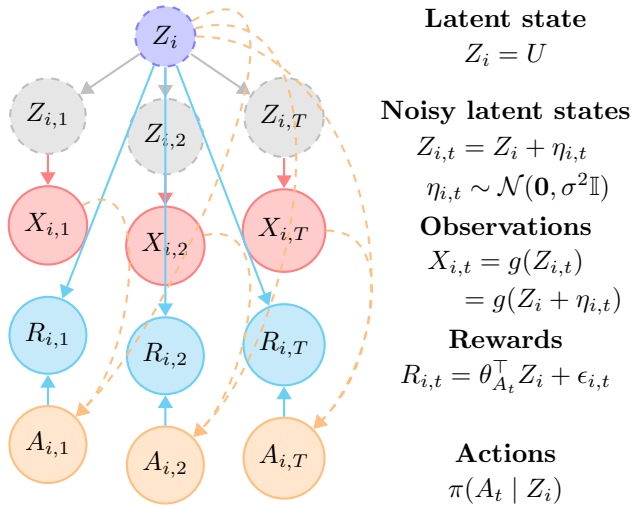


Figure 3.4: The structural causal model assumed in the ILB framework for an example instance i.

instance i. We provide two greedy algorithms, `CPG` and `FPG` for action selection. We theoretically demonstrate causal identifiability of the decision-making criteria under identifiable LVMs, and also provide empirical supporting results in a semi-synthetic decision-making environment e.g in (Figure 3.5), confirming that identifiable latent bandits are feasible to learn from data, albeit under specific assumptions.

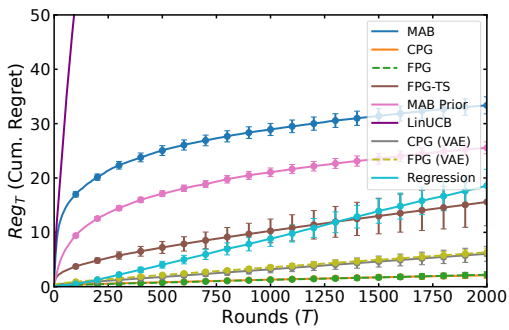


Figure 3.5: Cumulative regret results for ADCB (Kinyanjui & Johansson, 2022) comparing `CPG` and `FPG` in the identifiable latent bandit framework to baselines. Our results demonstrate the identifiability of the decision-making criteria under identifiable LVMs which can be learned from data albeit under specific assumptions.

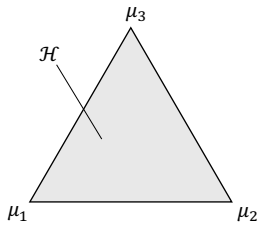
This exploration of identifiable latent bandits reveals several critical insights into the challenges and possibilities of learning latent variable models (LVMs) for optimal decision-making. A primary challenge lies in the inherent difficulty of learning identifiable LVMs, which necessitates specific identifying assumptions to ensure that the true underlying process can be recovered from observational data. When these assumptions are satisfied, it becomes feasible to learn the LVM, thereby enabling simple decision-making strategies to yield optimal results, as demonstrated in the identifiable latent bandit framework with with greedy strategies. However, these assumptions are not trivial to make; they impose specific conditions on the data-generating process, such as the structure of the latent state and the nature of the reward models, which may not always hold in real-world scenarios.

3.4 Challenge 3: Generalizing latent bandits to use looser latent structures of latent preference orderings

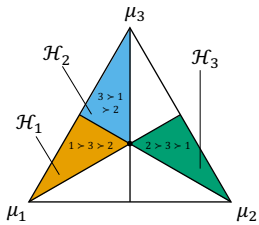
Another challenge that arises with latent bandits is that assuming availability of the full LVM can be too restrictive. This is because estimating a full LVM requires nontrivial assumptions as we illustrate with the previous challenge, it may not be identifiable from historical data, and may require a very large dataset even if it is. Moreover, requiring that all instances with state $Z = z$ follow the same reward distribution $p(R_a \mid Z = z)$ prevents instances from having individual *reward scales*: for example, two patients with a chronic condition could have the same subtype of disease z, which determines what therapies $a \in \mathcal{A}$ are preferred over which other therapies, but the two patients could have different tolerance for pain and give different ratings R_a for their symptoms under the same treatment a even if their relative preferences are the same. How would requiring a looser information structure of reward preference help to distinguish the true latent state from alternatives, and how could this benefit latent bandit decision making? To this end, Paper III (Mwai, Carlsson & Johansson, 2025) introduces latent bandits with latent state structure defined by *preference orderings* of actions.

Latent Preference Bandits (LPB) are a new latent bandit setting where each latent state $z \in \mathcal{Z} = [M]$ defines a preference ordering $O_z = (o_{z,1}, \ldots, o_{z,K})$ over actions, with rewards $R_a \sim \mathcal{N}(\mu_a, \sigma^2)$ satisfying $\mu_{o_{z,1}} \geq \cdots \geq \mu_{o_{z,K}}$. The LPB problem is illustrated in Figure 3.6 for the special case of reward means in the 2-dimensional simplex, compared to the standard multi-armed bandit (MAB) and the latent bandit problem from Hong et al. (2020a). Unlike traditional latent bandits, in LPB, two problem instances $(z, \boldsymbol{\mu}), (z', \boldsymbol{\mu}')$ with the same latent state $z = z'$ are guaranteed to have the same preference orderings but may not have the same distributions of rewards, which allows for modeling individual rating scales.

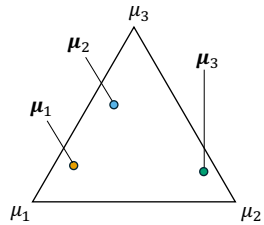
Towards an algorithm for the LPB problem, we propose the **lpbTS** regret minimization algorithm based on sampling from the posterior of the latent



(a) Multi-armed bandit. No constraints on reward means



(b) Latent preference bandit (this work). Known possible orderings, unknown reward means



(c) Latent bandit: Fully known possible reward means

Figure 3.6: Illustration of the latent preference bandit and related problems for reward means on the 2-simplex $\boldsymbol{\mu} \in \Delta^{K-1}$. In the MAB problem, no structure is known. In latent bandits, the full vector of reward means μ_z is known for each latent state z. In latent preference bandits, only the set of possible orderings is known (shown as colored segments), but two problem instances with the same latent state z may differ in their means as long as the orderings of their reward means are equal.

state and selecting the optimal arm for that state. With a history $\mathcal{D}_T = ((a_1, r_1), \ldots, (a_T, r_T))$ of the first T observations collected during exploration for a problem instance $(z, \boldsymbol{\mu})$, the likelihood of $\mathcal{D}_T$ under a state z with preference ordering O_z is

$$\mathcal{L}(\mathcal{D}_T \mid Z = z) = \prod_{t=1}^{T} p(r_t \mid a_t, z) = \int_{\boldsymbol{\mu} \in \mathcal{H}_z} p(\boldsymbol{\mu} \mid z) \prod_{t=1}^{T} p(r_t \mid a_t, \boldsymbol{\mu}, z)\, d\boldsymbol{\mu}.$$

This can be used to construct the posterior probability $p(Z = z \mid \mathcal{D}_T)$, provided that a well-specified parameter prior $p(\boldsymbol{\mu} \mid z)$ is known for each latent state z. In general, the constraint $\boldsymbol{\mu} \in \mathcal{H}_z$ means that no closed-form expression exists, and computing it exactly is intractable. As we aim to minimize the information needed about the latent variable, *we assume that no parameter prior is available.*

Without a parameter prior, the likelihood $p(r_t \mid a_t, z)$ is not fully defined, but we construct an upper bound on the likelihood by considering the mean configuration with the highest likelihood for the data restricted to the available orderings implied by $\mathcal{Z}$. For all states z,

$$\mathcal{L}(\mathcal{D}_T \mid Z = z) \leq \sup_{\boldsymbol{\mu} \in \mathcal{H}_z} \prod_{t=1}^{T} p(r_t \mid a_t, \mu_{a_t}).$$

With Gaussian rewards, maximizing this *upper* bound corresponds to minimizing the mean squared error of $\boldsymbol{\mu}$ in predicting the observed reward, constrained to the set $\mathcal{H}_z$. Thus, under the assumption that z is the correct latent state,

we may estimate the mean parameters as follows:

$$\hat{\boldsymbol{\mu}}_z \coloneqq \arg\min_{\boldsymbol{\mu}\in\mathcal{H}_z} -\ell(\mathcal{D}_T \mid \boldsymbol{\mu}), \quad \text{where} \quad -\ell(\mathcal{D}_T \mid \boldsymbol{\mu}) \propto \sum_{t=1}^{T} \frac{(r_t - \mu_{a_t})^2}{\sigma^2}. \tag{3.3}$$

With $\{\hat{\boldsymbol{\mu}}_z\}$ the minimizers of (3.3) for all z, we construct an *optimistic* posterior estimate,

$$\forall z : \hat{p}(z \mid \mathcal{D}_t) \coloneqq \frac{1}{\alpha} p(\mathcal{D}_t \mid \hat{\boldsymbol{\mu}}_z), \tag{3.4}$$

where α is the normalization constant.

The **lpbTS** algorithm selects the optimal arm for a state sampled from the approximate posterior (3.4). The constrained maximum-likelihood estimation (MLE) problem in (3.3), solved for each state, is a quadratic program with linear inequality constraints that, as we show with Proposition 3, can be solved using off-the-shelf solvers for isotonic regression (Barlow & Brunk, 1972).

Proposition 3 *Let* $n_a = \sum_{t=1}^{T} \mathbb{1}[a_t = a]$ *and define* $w_a = \frac{n_a}{\sigma_a^2}$. *Next, let* $O_z = (o_1, \ldots, o_K)$ *be the preference ordering of latent state* z. *Then, the solution to the isotonic regression problem with outcomes* $y_a = \frac{1}{n_a} \sum_{t:a_t=a} r_t$ *and sample weights* w_a

$$\textit{minimize}_{\boldsymbol{\mu}\in\mathbb{R}^d} \sum_{a=1}^{K} w_a(\mu_a - y_a)^2 \quad \textit{subject to} \quad \mu_{o_K} \le \mu_{o_{K-1}} \le \cdots \le \mu_{o_1}$$

solves the constrained MLE problem in (3.3).

Empirically, we demonstrate (Figure 3.7) that the LPB problem is solvable, and that **lpbTS** is comparable in performance to **mTS** (Hong et al., 2020a) when instance reward means have a fixed reward scale in the latent states, and that adding the ordering constraints O is vastly beneficial compared to no structure. We also demonstrate the benefit of using a more general latent structure O compared to a latent mean vector of rewards, where **lpbTS** outperforms **mTS** with differing individual reward scales — because a latent model comprising mean vectors is misspecified when absolute reward scales can vary for different latent state instances. This is also demonstrated on real-world datasets, the MovieLens (Harper & Konstan, 2015) datasets where actions represent movie choices, rewards are ratings of movies, and latent states are groups of users (Figure 3.8).

Beyond algorithm design to leverage the LPB structure, we are also able to understand latent bandits more foundationally via empirical investigation, and connect our understanding to why the latent preference ordering O is beneficial for sample complexity. We realize that the $O(\sqrt{MT \log T})$ upper bound for latent bandit algorithms like **mTS** (Hong et al., 2020a) does not really explain the latent structure, and it can be achieved simply by restricting the action set (Proposition 4).

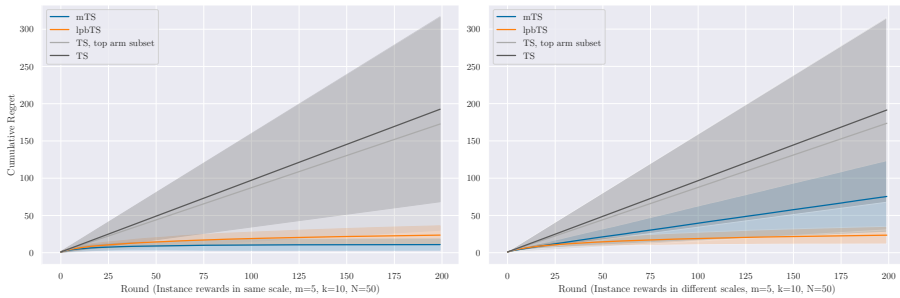
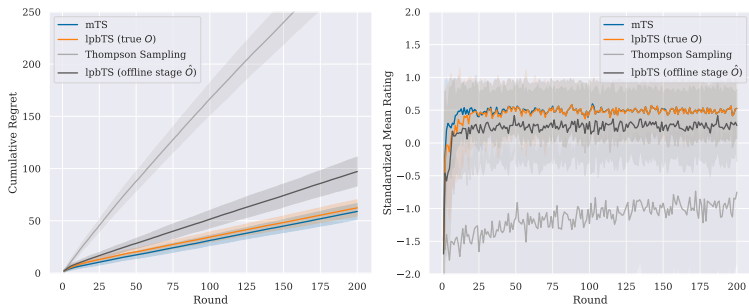
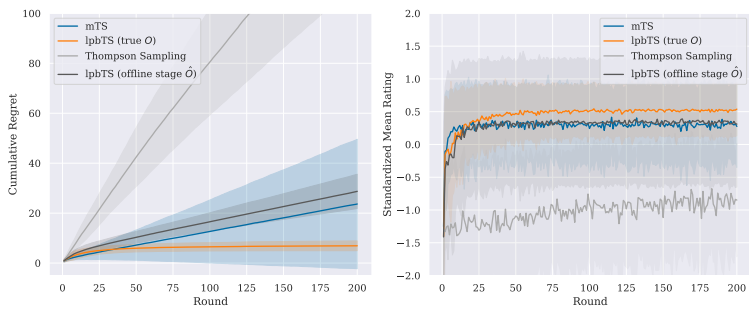


Figure 3.7: **lpbTS** is comparable to latent bandit baselines when instance rewards lie in the same scale (***Left***) and outperforms baselines when instance rewards lie in different reward scales (***Right***).



(a) Movie ratings in the same scale for users in a latent state



(b) Movie ratings in different scales for users in a latent state

Figure 3.8: MovieLens Experiment, 20M Dataset. Results match theory: **lpbTS** is comparable to mTS in (a), outperforms in (b), and the two-stage recovery of O is empirically validated.

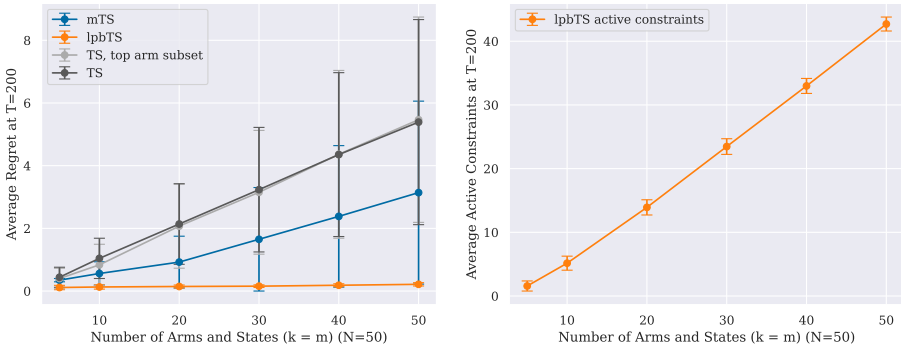


Figure 3.9: Varying the number of arms K, and $M = K$ ($N = 50, T = 200, K \in [5, 10, 20, 30, 40, 50]$). ***Left:*** Observed average regret at $T = 200$. ***Right:*** Observed average active constraints.

Proposition 4 *Consider the following algorithm. Whenever $K < M$, restrict the action set to the subset $\mathcal{A}^*_{\mathcal{Z}}$ of optimal $u < M$ arms of which each is optimal in at least one latent state, $\mathcal{A}^*_{\mathcal{Z}} = \{a \in [K] : \exists z \in \mathcal{Z}$ such that $a^*_z = a\}$, and run a standard MAB algorithm restricted to $\mathcal{A}^*_{\mathcal{Z}}$. When $K \geq M$, run a standard MAB algorithm on $\mathcal{A} = [K]$. This procedure achieves $O(\sqrt{\min(K, M)T})$ regret in the worst case on the latent bandit and latent preference bandit problems.*

By investigating how the active constraints (in isotonic regression) in **lpbTS** change when K and M vary, we are able to understand the latent preference bandit problem better and generalize this for the latent bandit problem. For example, when $M = K$ and K increases, we observe that (Figure 3.9) the number of active constraints grows. This is because $M = O(K)$, but the number of possible permutations grows like $K!$, so the probability of having large differences between states grows when $M = K$ and K grows. This is not predicted by an $O(\sqrt{\min(K, M)T})$ bound since $M = K$. It is explained by the fact that the true latent state stands out more with high probability, and the empirical isotonic means $\hat{\mu}_z$ become less likely to align with the neighboring states (the most confusable states) relative to the true state, resulting in a higher number of active constraints.

Chapter 4

Understanding Bandits with Switching Constraints in Fixed-confidence Pure Exploration

In real-world treatment personalisation settings, such as in chronic disease treatment, bandit algorithms must often contend with practical constraints beyond sample efficiency. A critical constraint is the cost or limitation on switching between treatments. Switching treatments has costs for the patient because every time a treatment is changed, the patient has to weave off their current therapy and get used to the new treatment and its potential side effects. This chapter explores how exploration in bandit problems adapts to **switching constraints**, where the number of action switches is restricted, while exploring towards an optimal action, and how structuring exploration can help to yield solutions. The focus is on the fixed-confidence pure exploration (FC-PE) setting, continuing the theme of understanding exploration from the previous chapter (Section 3.2).

With the total number of arm switches S_τ as the number of successive plays where the arms differ, $S_\tau = \sum_{t=2}^{\tau} [a_t \neq a_{t-1}]$, the goal is to design a search strategy ϕ to:

Minimize the expected number of arm plays τ required to identify an optimal arm with confidence at least $1 - \delta$ for a given $\delta > 0$, while limiting the expected rate of switching arms to $\alpha \in [0, 1]$.

$$\begin{aligned} \underset{\phi}{\text{minimize}} \quad & \mathbb{E}_\phi[\tau] \\ \text{subject to} \quad & \mathbb{P}(\mu_{\hat{a}_\tau} < \mu^*) \leq \delta \\ & \mathbb{E}_\phi[S_\tau] \leq \alpha \mathbb{E}_\phi[\tau] \end{aligned} \tag{4.1}$$

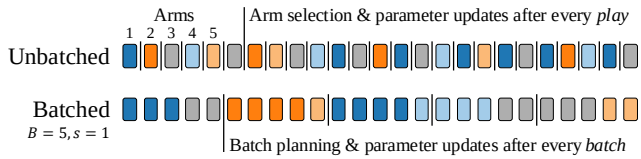


Figure 4.1: Illustration of batched arm plays used to limit the arm switching frequency in a 5-arm problem. The number of plays of each arm is the same.

This formulation, however, poses a challenge: the switching rate constraint depends on the expected stopping time $\mathbb{E}_\phi[\tau]$, which is unknown during execution.

Paper IV (Mwai, Malekipirbazari & Johansson, 2025) resolves the challenge in objective 4.1 as a reformulation with a *batched* variant of the problem (Figure 4.1) that provides a practical and well-defined alternative to the original constraint. With batched exploration using batches of size B assumed fixed and known, the number of switches in exploration are attributed either to: switching between arms *within* the batches when changing from one successive arm play segment to the next, or to changing arms *between* batches. The goal is re-formulated to be to:

Minimize the expected number of batches β required to identify an optimal arm, with confidence at least $1-\delta$, while limiting the arm switches within the batch to be at most $s \in \{0, ..., \min(K-1, B-1)\}$,

$$\begin{aligned} \underset{\phi}{\text{minimize}} \quad & \mathbb{E}_\phi[\beta] \\ \text{subject to} \quad & \mathbb{P}\left(\mu_{\hat{a}_\beta} < \mu^*\right) \leq \delta \\ & S^b \leq s, \ \forall b \in \mathbb{N} \end{aligned} \tag{4.2}$$

where S^b is the number of switches in batch b, and $\hat{a}_\beta$ is the recommended arm after β batches.

Given that the batch size is assumed fixed and known, we can index all possible *sparse batch configurations* c of integer arm plays in a batch that satisfy the desired switching limit. For a given number of arms K, batch size B and switching limit s, we denote this set $\mathcal{C}_{B,s}^K$,

$$\mathcal{C}_{B,s}^K := \left\{ c \in \mathbb{N}^K : \sum_{a=1}^{K} c_a = B, \ \ \|c\|_0 \leq s+1 \right\} . \tag{4.3}$$

where $\|\cdot\|_0$ denotes the ℓ_0-norm, which counts the number of nonzero elements in the vector, $\|x\|_0 := \sum_{i=1}^K \mathbb{1}[x_i \neq 0]$. Each element $c = [c_1, ..., c_K]^\top \in \mathcal{C}_{B,s}^K$ represents a configuration that can be executed in a single batch and each coordinate c_a represents the number of times arm a will be played in the batch. c is *sparse* if there are arms a such that $c_a = 0$. Through $\mathcal{C}_{B,s}^K$, we state a lower bound for batch-playing bandits that obey the switching constraint:

Theorem 1 *Let $\Sigma^{\mathcal{C}} := \Sigma^{|\mathcal{C}^K_{B,s}|-1}$ be the simplex over batch configurations of size B that use fewer than s switches. Given a confidence level $\delta \in (0,1)$, for any algorithm that returns the best arm with probability at least $1-\delta$, and for any bandit problem $\mu \in \mathbb{R}^K$, the following holds:*

$$\mathbb{E}_\mu[\beta] \geq T^*_{bc}(\mu) \cdot \mathrm{kl}(\delta, 1-\delta), \tag{4.4}$$

where the characteristic time $T^*_{bc}(\mu)$ is given by

$$T^*_{bc}(\mu)^{-1} := \sup_{p \in \Sigma^{\mathcal{C}}} \inf_{\lambda \in \mathrm{Alt}(\mu)} \sum_{a=1}^{K} \sum_{c \in \mathcal{C}^K_{B,s}} p_c c_a d(\mu_a, \lambda_a). \tag{4.5}$$

In (4.5), the supremum is computed over the possible probability distributions over sparse configurations, thereby incorporating the in-batch switching limit into the batch play optimization.

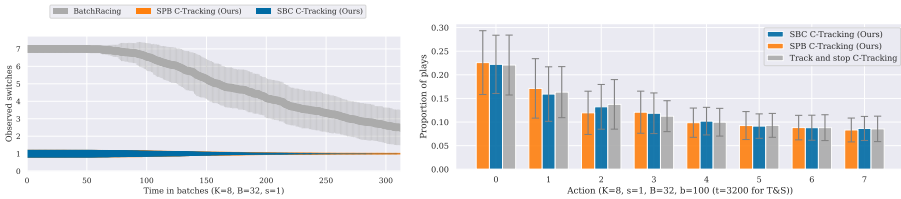
Towards algorithms for this setting, we take inspiration from Garivier and Kaufmann (2016) *track-and-stop* algorithm design strategy, which aims to *track* the optimal arm playing proportions $w^*(\hat{\mu})$ of the lower bound in (2.4),

$$w^*(\hat{\mu}) := \arg\max_{w \in \Sigma^K} \inf_{\lambda \in \mathrm{Alt}(\hat{\mu})} \left(\sum_{a=1}^{K} w_a d(\hat{\mu}_a, \lambda_a) \right) . \tag{4.6}$$

However, applying the track-and-stop framework in our setting requires imposing a switching constraint in the tracking rule. We cannot impose sparsity in the tracked proportions w^* without destroying the solution to (4.6). If an arm a is never played, $w_a = 0$, the adversary λ can exploit this and differ arbitrarily for that arm, rendering the lower bound infinite. This is also evident from Lemma 4 in Garivier and Kaufmann (2016) which would be violated if $\exists a : w^*_a = 0$. Neither is it a good idea to play configurations to track the proportions p^* that solve (4.5). The solution is not necessarily unique and, even if it is, the number of possible configurations is exponential, making exploring (tracking) all of them infeasible. Moreover, the number of batches where a configuration is played is not itself of interest, only that the resulting distribution of arm plays is optimal. Instead, we track the optimal arm proportions with suitably chosen batches after making an observation (Observation 1):

Observation 1 *If the optimal arm allocation w^* in (4.6) is "realizable" under $\mathcal{C}$ (with $\mathcal{C} = \mathcal{C}^K_{B,s}$), i.e., $\exists p^* \in \Sigma^{\mathcal{C}}$ such that $\sum_{c \in \mathcal{C}} p^*_c c = w^*(\hat{\mu})$, then p^* are minimizers of (4.5).*

With this, we present a batch selection objective (selection rule) for this setting to find a feasible batch $\tilde{c}$:



(a) Observed switches and stopping, along time in batches for SBC and SPB C-Tracking (Ours) with $s = 1, B = 32$ vs BatchRacing (Baseline).

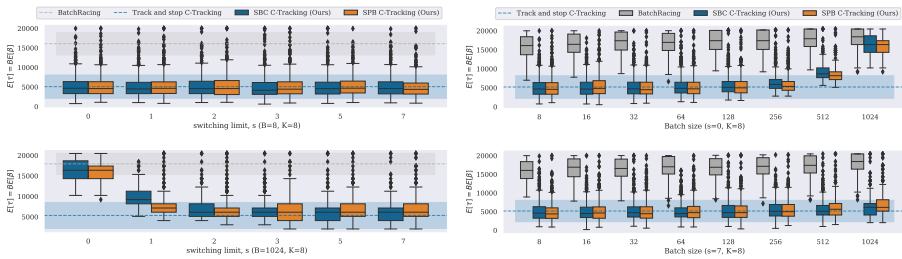
(b) Proportions of arm plays for SBC and SPB (Ours) after 100 batches (3200 plays, with $B = 32$, $s = 1$) and C-Tracking after 3200 plays.

Figure 4.2: SBC and SPB stop quicker even with a restrictive switching limit and match well to the *optimal, unbatched* Track-and-Stop C-tracking baseline in tracking proportions.

$$\tilde{c} \in \arg\min_{c \in \mathcal{C}} \sum_{a=1}^{K} \Big(d_a(b) - c_a \Big)_+ \tag{4.7}$$

where c_a are the number of plays of arm a in the batch configuration c, $d_a(b) := \bar{w}_a(b) - N_a(b)$ is the deficit for arm a in batch b. Here, $\bar{w}(b) = B \sum_{i=0}^{b-1} w^{\epsilon_i}(\hat{\mu}_i)$ are the C-tracking (Garivier & Kaufmann, 2016), goal proportions with $w^{\epsilon}(\hat{\mu})$ the L_∞-projection of $w^*(\hat{\mu})$ in (4.6) onto $\Sigma_\epsilon^K = \{w \in \mathbb{R}_+ : \sum_a w_a = 1, \min_a w_a \geq \epsilon\}$. We aim to minimize the total positive deficit $D(b) := \sum_{a=1}^{K} (d_a(b))_+$, where $(x)_+ = \mathbb{1}[x > 0]x$. Unlike the lower bound problem (4.5), (4.7) can actually be solved in polynomial time through a greedy algorithm. However, the solution is not unique. For example, if more than $s + 1$ arms have positive deficit, there are cases where the allocations to the selected arms in the batch can be decided partially arbitrarily. Once the deficit of selected arms has been removed, the choice of how to distribute remaining plays between them won't alter (4.7).

We present two algorithms: i) *Sparse Batch Configurations* (SBC) C-Tracking SBC a greedy batch configuration construction algorithm that puts the remaining allocation on the arm with the largest remaining fractional deficit. ii) *Sparse-Projected Batch (SPB)* C-tracking algorithm where the allocations in the batch are distributed proportionally to the deficits of the selected arms. The idea in SPB is to project the normalized positive deficits between expected and actual plays $(\bar{d}(b))_+ = \frac{(d(b))_+}{\sum_{a \in \mathcal{A}} (d_a(b))_+}$ onto an *(s+1)-sparse simplex* and the batch configuration is constructed according to the resulting sparse proportions. By proving that batch configurations selected according to (4.7) track the optimal arm proportions in (4.6), we also show that both SBC and SPB C-Tracking match the lower bound in (2.4) in the high-certainty limit.



(a) Comparison of stopping times over switching limits $s \in \{0, 1, 2, 3, 5, 7\}$ in SBC and SPB C-Tracking, and BatchRacing, with batch sizes $B \in \{8, 1024\}$. Track-and-stop C-Tracking is not batched.

(b) Effect of batch size on the stopping times for SBC and SPB C-Tracking ($s \in \{0, 7\}$), and BatchRacing, with $B \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$.

Figure 4.3: To balance generality of the abstraction through $\mathcal{C}^K_{B,s}$, and practical insight, we provide simulation results showing that SBC and SPB C-Tracking perform well except under extreme conditions, specifically when the batch size is large and the switching constraint is stringent, due to wasted plays.

Chapter 5

Semi-synthetic Causal Benchmark for Evaluating Treatment Personalization algorithms

The success of reinforcement learning as a sequential decision-making paradigm has been greatly facilitated by the availability of standard benchmark problems which enable researchers to develop, test, and compare reinforcement learning algorithms (Kuo et al., 2022). In many healthcare systems, there is plenty of data collected in electronic health records (EHRs) (Ambinder, 2005) that could be valuable if leveraged to design sequential decision-making systems to improve healthcare. However, evaluating algorithms in an *online* setting, where actions directly affect patients, is often infeasible due to ethical and safety constraints, even when extensive real-world data is available. We cannot experimentally manipulate treatments or run exploratory policies on patients to gather data, which makes simulators essential for iterative algorithm development and benchmarking.

In addition to the online treatment setting constraints, challenges of accessibility attributable to justifiable privacy concerns regarding disclosure of private patient information, accessibility remains a challenge. In spite of this challenge, several databases containing longitudinal data are publicly available, for example the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, containing longitudinal data on Alzheimer's disease (AD) patients and cognitively normal controls. Another, the MIMIC-III ("Medical Information Mart for Intensive Care") (Johnson et al., 2016) is a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital. When applicable, researchers have widely used such datasets in their empirical studies. However, even when available, the datasets are small, whereas sequential decision-making techniques usually require a large number of training samples (Yu et al., 2021).

Researchers have therefore resorted to building synthetic benchmark simulators which have many advantages but often lack the intricacies observed in reality (Hernán, 2019). For simulators to be useful, preserving causal relationships between clinical variables is critical. Without causal fidelity, algorithm evaluations may be misleading, particularly in sequential treatment settings where decision outcomes depend on dynamic, interrelated factors. Yet, realistic benchmarks that combine causal realism with healthcare complexity remain scarce. While simulators like IHDP (Hill, 2011) and ACIC (Dorie et al., 2019) have been valuable for causal effect estimation, they rely on simplified, static mathematical response surfaces and are not designed for sequential decision-making as in bandit settings. More data-driven approaches (Chan et al., 2021; Neal, Huang & Raghupathi, 2020; Kuo et al., 2022) increase realism but often overlook underlying causal mechanisms. This underscores the need for hybrid benchmarks that integrate real clinical data with domain-expert causal knowledge (Hernán, 2019), enabling both realism and validity in evaluating bandit algorithms in personalized medicine. This poses a challenge: Is it possible to design an environment for evaluating sequential decision-making algorithms with realistic healthcare data that matches clinical statistics in EHRs and a causal structure of the generating process from domain knowledge?

Paper V (Kinyanjui & Johansson, 2022), introduced a method for designing a semi-synthetic benchmark simulator for longitudinal Alzheimer's disease data that incorporates verifiable causal domain knowledge. The Alzheimer's Disease Causal estimation Benchmark (ADCB) was designed. The simulator was fit to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and ground-crafted components incorporating results from comparative treatment trials and observational treatment patterns. Tuning parameters were also incorporated, which causally alter the nature and difficulty of the learning tasks, such as latent variables, effect heterogeneity, length of observed subject history, behaviour policy and sample size. Moreover, ADCB also generates longitudinal data that includes potential outcomes for all treatments at each step in the longitudinal axis.

The design started by positing a causal graph for the variables of interest at the baseline time point of observation based either on models fit to the ADNI data, on hand-crafted functions or on results from AD literature. This causal graph is shown in Figure 5.1.

A usage example of using the ADCB simulator to compare standard estimators of causal effects was outlined in the work, where a) a single time point is used to estimate average and personalized treatment effects, and b) a time series of patient history is used (Figure 5.2(a) and Figure5.2(b)).

To make the ADCB simulator a more robust environment for studying latent bandits, the latent states have since been further expanded from two latent states to six, and also continuous latent states have been incorporated. In addition a gym environment has since been developed with logged data from the ADCB simulator, where bandit algorithms can be compared.

For bandit algorithms, ADCB provides a powerful testing ground by simulating longitudinal, high-dimensional data with verifiable causal relationships, and it was used for the experimental study in Paper I and Paper II. For example, in

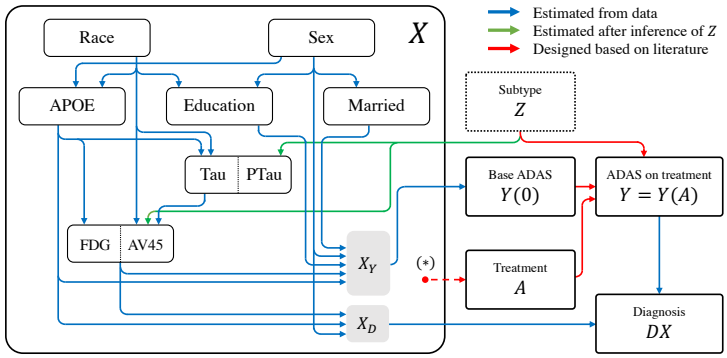
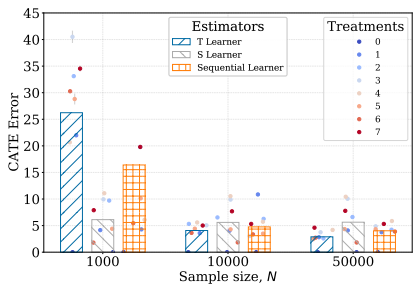
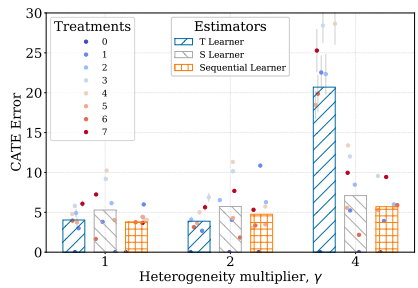


Figure 5.1: Assumed causal graph for the ADCB simulator. Arrows indicate causal dependencies, with colour representing how the mechanism was determined. Blue dependencies were completely estimated from data, green were fit once the subtype Z was inferred, and red were designed based on the Alzheimer's disease literature.



(a) CATE mean squared error varying with sample size, N. ϵ=0.1, γ=2, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$



(b) CATE error varying with heterogeneity, γ. ϵ=0.1, Sample size, $N = 10,000$, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$

Figure 5.2: A usage example of using the ADCB simulator to compare standard estimators of causal effects

Paper I, latent bandit algorithms were evaluated on ADCB data in the ADCB bandit gym environment with $K = 8$ actions and $Z = 6$ latent states, where the simulator generated outcomes Y_t as $Y_t(A, X, Z) = \Phi(X, Z) + \Delta(A_t, Z) + \xi$, where Φ models untreated cognitive function (fit to real ADNI data), Δ reflects treatment effects moderated by latent state Z, and $\xi \sim \mathcal{N}(0, \sigma^2)$ adds noise. Bandit algorithms aimed to identify the optimal action $a^* = \arg\max_a \mathbb{E}[Y_t \mid A = a, Z]$, requiring exploration of latent states inferred from observed contexts X. ADCB's flexibility such as ability to tweak assumptions like introducing confounding or varying effect heterogeneity further enhances its utility. The flexibility is also in it's system design, as demonstrated in Paper II, where the simulator environment was further adapted to include mixed latent states (categorical and continuous), while still being causally grounded.

Ultimately, the work on ADCB demonstrates that simulators grounded in domain knowledge and real data are essential for advancing sequential decision-making (e.g. with bandit algorithms) in healthcare. By providing access to counterfactual outcomes and tunable parameters, it bridges the gap between theoretical benchmarks and practical challenges, ensuring practical design of algorithms that are clinically relevant.

Chapter 6

Conclusion

This thesis has advanced the understanding and application of multi-armed bandits (MABs) to the domain of personalized medicine, with a particular emphasis on chronic diseases such as Alzheimer's Disease (AD). By addressing the critical challenge of optimizing the exploration-exploitation trade-off in treatment personalization, the work in this thesis has developed novel strategies that enhance sample efficiency and accommodate practical clinical constraints, complemented with foundational insights. These contributions are pertinent in personalized medicine where long exploration is impractical due to the high costs associated with patient well-being and medical resources. The primary contributions of this research are threefold, each addressing distinct challenges in applying MABs to personalized medicine.

In Chapter 3, results from studying the Latent Bandit framework in the fixed-confidence pure exploration (FC-PE) setting were presented, providing foundational insights into how latent structures enhance efficiency to significantly reduce the number of trials required to identify optimal treatments. The Identifiable Latent Bandit (ILB) framework, also presented in Chapter 3, tackled the challenge of learning reward models from observational historical data, proving that identifiable LVMs can be learned offline with historical data, and that integrating offline LVM learning with online decision-making provably leads to optimal decision-making, though this relies on stringent assumptions about the data-generating process. Moreover, the Latent Preference Bandits (LPB) framework also presented in Chapter 3 extended latent bandits to incorporate preference orderings rather than fixed reward distribution vectors, allowing for individual variations in reward scales. The proposed **lpbTS** algorithm for LPB demonstrated comparable or superior performance to traditional latent bandits, particularly when reward scales differ across instances. This generalization highlights the potential of looser structural priors to improve adaptability in personalized medicine.

Chapter 4 reformulated the FC-PE problem to address the practical constraint of limiting treatment switches, crucial for reducing patient burden. By structuring exploration with a batched approach, the Sparse Batch Configurations (SBC) and Sparse-Projected Batch (SPB) C-Tracking algorithms

effectively minimized the expected number of batches while respecting switching limits. Theoretical and empirical analyses showed that these algorithms are optimal in performance, except under extreme conditions of large batch sizes and stringent constraints.

Chapter 5 presented the ADCB simulator, a semi-synthetic benchmark that combines real-world ADNI data with domain-informed causal structures, and AD therapies and treatment policies from literature. The simulator provides a robust platform for evaluating bandit algorithms, offering tunable parameters and counterfactual outcomes that mirror clinical complexities. Its utility was demonstrated across multiple chapters, underscoring its value as a tool for bridging theoretical and practical research.

While specific limitations are detailed in the appended papers, several limitations in the approaches merit highlighting. In the latent bandits work, the assumption of stationary latent states, where patient subtypes remain fixed over time overlooks dynamic disease progression. This could potentially lead to suboptimal long-term personalization, and it therefore necessitates non-stationary extensions. Also, the Identifiable Latent Bandit framework relies on strong identifiability and learnability assumptions, which may not hold in noisy, confounded real-world electronic health records (EHRs), risking biases in the presence of unmeasured confounders. For batched bandits with switching constraints, the fixed, known batch size overlooks variable clinical cycles (e.g., influenced by adherence or side effects), which could lead to inefficient exploration if batches do not align with real horizons, so adaptive batching could enhance robustness. These limitations highlight a broader tension: while structured assumptions enable tractable solutions, they may trade off generalizability in complex, non-stationary healthcare environments, necessitating hybrid approaches that balance methodological rigor with flexibility.

As directions for future work, the LPB framework stands out as fertile ground, given the flexibility promised with preference orderings. Extending it to a contextual bandit version could yield more refined personalization, especially in high-dimensional settings like genomics-based treatment personalization. Furthermore, with the rise of large language models (LLMs) and preference-based learning such as in reinforcement learning from human feedback (RLHF), exploring dueling bandits with latent structures could be interesting to investigate if leveraging latent structures could yield faster convergence in dueling bandits. More broadly, the insights from this thesis, particularly on leveraging historical data and exploration under constraints transfer naturally to standard reinforcement learning (RL) in clinical contexts, starting with best policy identification in Markov decision processes, which is inspired by best arm identification studied in this thesis.

Ultimately, beyond refining methods, fostering interdisciplinary collaborations with clinicians promises the greatest leap for personalized medicine research in creating ethical frameworks to blend algorithmic tools with clinical expertise. This synergy will yield trustworthy, adaptive, interactive systems prioritizing patient-centered outcomes, in an era of data-driven medicine.

Bibliography

Blennow, K., de Leon, M. J., & Zetterberg, H. (2006). Alzheimer's disease. *The lancet*, *368*(9533), 387–403 (cit. on p. 3).

Alzheimer's Association. (2024). 2024 alzheimer's disease facts and figures. *Alzheimer's Dementia*, *20*(5), 3708–3821. https://doi.org/10.1002/alz.13809 (cit. on p. 3).

Aletaha, D., & Smolen, J. S. (2018). Diagnosis and management of rheumatoid arthritis: A review. *Jama*, *320*(13), 1360–1372 (cit. on p. 3).

Fraenkel, L., Bathon, J. M., England, B. R., St. Clair, E. W., Arayssi, T., Carandang, K., Deane, K. D., Genovese, M., Huston, K. K., Kerr, G., et al. (2021). 2021 american college of rheumatology guideline for the treatment of rheumatoid arthritis. *Arthritis & Rheumatology*, *73*(7), 1108–1123 (cit. on p. 3).

Raharja, A., Mahil, S. K., & Barker, J. N. (2021). Psoriasis: A brief overview. *Clinical Medicine*, *21*(3), 170–173 (cit. on p. 3).

Kim, W. B., Jerome, D., & Yeung, J. (2017). Diagnosis and management of psoriasis. *Canadian Family Physician*, *63*(4), 278–285 (cit. on p. 3).

Singh, J. A., Saag, K. G., Bridges Jr, S. L., Akl, E. A., Bannuru, R. R., Sullivan, M. C., Vaysbrot, E., McNaughton, C., Osani, M., Shmerling, R. H., et al. (2016). 2015 american college of rheumatology guideline for the treatment of rheumatoid arthritis. *Arthritis & rheumatology*, *68*(1), 1–26 (cit. on p. 3).

Murphy, S. A., Collins, L. M., & Rush, A. J. (2007). Customizing treatment to the patient: Adaptive treatment strategies. (Cit. on p. 3).

Farlow, M. R., Miller, M. L., & Pejovic, V. (2008). Treatment options in alzheimer's disease: Maximizing benefit, managing expectations. *Dementia and geriatric cognitive disorders*, *25*(5), 408–422 (cit. on p. 3).

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., Ballard, C., Banerjee, S., Burns, A., Cohen-Mansfield, J., et al. (2017). Dementia prevention, intervention, and care. *The lancet*, *390*(10113), 2673–2734 (cit. on p. 3).

Grossberg, G. T., Tong, G., Burke, A. D., & Tariot, P. N. (2019). Present algorithms and future treatments for alzheimer's disease. *Journal of Alzheimer's Disease*, *67*(4), 1157–1171 (cit. on p. 3).

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294 (cit. on pp. 4, 7, 9).

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, *30*(3), 755–770 (cit. on pp. 4, 7).

Chernoff, H. Sequential models for clinical trials. In: In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 4: Biology and problems of health. 5.* University of California Press. 1967, 805–813 (cit. on pp. 4, 7).

Gittens, J., & Dempster, M. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B: Methodological*, *41*, 148–177 (cit. on pp. 4, 7).

Lai, T., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*, 4–22 (cit. on pp. 4, 7).

Li, L., Chu, W., Langford, J., & Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In: In *Proceedings of the 19th international conference on world wide web.* 2010, 661–670 (cit. on pp. 4, 7, 12, 13).

Chapelle, O., & Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, *24* (cit. on pp. 4, 10, 12).

Bouneffouf, D., Rish, I., & Aggarwal, C. Survey on applications of multi-armed and contextual bandits. In: In *2020 ieee congress on evolutionary computation (cec).* 2020, 1–8. https://doi.org/10.1109/CEC48606.2020.9185782 (cit. on pp. 4, 12).

Yancey, K. P., & Settles, B. A sleeping, recovering bandit algorithm for optimizing recurring notifications. In: In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining.* 2020, 3008–3016 (cit. on pp. 4, 12).

O'Brien, C., Wu, H., Zhai, S., Guo, D., Shi, W., & Hunt, J. J. (2022). Should i send this notification? optimizing push notifications decision making by modeling the future. *arXiv preprint arXiv:2202.08812* (cit. on pp. 4, 12).

Chu, W., Li, L., Reyzin, L., & Schapire, R. Contextual bandits with linear payoff functions. In: In *Proceedings of the fourteenth international conference on artificial intelligence and statistics.* JMLR Workshop and Conference Proceedings. 2011, 208–214 (cit. on pp. 4, 12, 13).

Agrawal, S., & Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In: In *International conference on machine learning.* PMLR. 2013, 127–135 (cit. on pp. 4, 12).

Zhou, L. (2015). A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326* (cit. on pp. 4, 12, 13).

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, *24* (cit. on pp. 4, 13).

Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901* (cit. on p. 4).

Riachi, E., Mamdani, M., Fralick, M., & Rudzicz, F. (2021). Challenges for reinforcement learning in healthcare. *arXiv preprint arXiv:2103.05612* (cit. on p. 4).

Zhang, C., Agarwal, A., Daumé III, H., Langford, J., & Negahban, S. N. (2019). Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301* (cit. on pp. 4, 16).

Oetomo, B., Perera, R. M., Borovica-Gajic, R., & Rubinstein, B. I. (2023). Cutting to the chase with warm-start contextual bandits. *Knowledge and Information Systems*, *65*(9), 3533–3565 (cit. on pp. 4, 16).

Bui, L., Johari, R., & Mannor, S. (2012). Clustered bandits. *arXiv preprint arXiv:1206.4169* (cit. on pp. 4, 16).

Bouneffouf, D., Parthasarathy, S., Samulowitz, H., & Wistub, M. (2019). Optimal exploitation of clustering and history information in multi-armed bandit. *arXiv preprint arXiv:1906.03979* (cit. on pp. 4, 16).

Maillard, O.-A., & Mannor, S. Latent bandits. In: In *International conference on machine learning*. PMLR. 2014, 136–144 (cit. on pp. 4, 16).

Hong, J., Kveton, B., Zaheer, M., Chow, Y., Ahmed, A., & Boutilier, C. Latent bandits revisited (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin, Eds.). In: *Advances in neural information processing systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin, Eds.). Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. *33*. Curran Associates, Inc., 2020, 13423–13433 (cit. on pp. 4, 16, 17, 20, 22, 24).

Zhou, L., & Brunskill, E. (2016). Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743* (cit. on pp. 4, 16).

Hong, J., Kveton, B., Zaheer, M., Chow, Y., Ahmed, A., Ghavamzadeh, M., & Boutilier, C. (2020b). Non-stationary latent bandits. *CoRR*, *abs/2012.00386* (cit. on pp. 4, 16).

Galozy, A., & Nowaczyk, S. (2023). Information-gathering in latent bandits. *Knowledge-Based Systems*, *260*, 110099 (cit. on p. 4).

Garivier, A., & Kaufmann, E. Optimal best arm identification with fixed confidence (V. Feldman, A. Rakhlin & O. Shamir, Eds.). In: *29th annual conference on learning theory* (V. Feldman, A. Rakhlin & O. Shamir, Eds.). Ed. by Feldman, V., Rakhlin, A., & Shamir, O. *49*. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, 2016, 998–1027 (cit. on pp. 4, 11, 12, 17, 19, 29, 30).

Kaufmann, E. (2020). *Contributions to the optimal solution of several bandit problems* [Doctoral dissertation, Université de Lille]. (Cit. on pp. 4, 11).

Agrawal, S., Juneja, S., Shanmugam, K., & Suggala, A. S. (2023). Optimal best-arm identification in bandits with access to offline data. *arXiv preprint arXiv:2306.09048* (cit. on p. 4).

Arora, R., Dekel, O., & Tewari, A. (2012). Online bandit learning against an adaptive adversary: From regret to policy regret. *arXiv preprint arXiv:1206.6400* (cit. on p. 4).

Dekel, O., Ding, J., Koren, T., & Peres, Y. Bandits with switching costs: T 2/3 regret. In: In *Proceedings of the forty-sixth annual acm symposium on theory of computing*. 2014, 459–467 (cit. on p. 4).

Rouyer, C., Seldin, Y., & Cesa-Bianchi, N. An algorithm for stochastic and adversarial bandits with switching costs. In: In *International conference on machine learning*. PMLR. 2021, 9127–9135 (cit. on p. 4).

Amir, I., Azov, G., Koren, T., & Livni, R. (2022). Better best of both worlds bounds for bandits with switching costs. *Advances in neural information processing systems*, *35*, 15800–15810 (cit. on p. 4).

Li, Y., Preiss, J. A., Li, N., Lin, Y., Wierman, A., & Shamma, J. S. Online switching control with stability and regret guarantees. In: In *Learning for dynamics and control conference*. PMLR. 2023, 1138–1151 (cit. on p. 4).

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240 (cit. on pp. 5, 34).

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43–68 (cit. on pp. 5, 34).

Chan, A. J., Bica, I., Huyuk, A., Jarrett, D., & van der Schaar, M. (2021). The medkit-learn (ing) environment: Medical decision modelling through simulation. *arXiv preprint arXiv:2106.04240* (cit. on pp. 5, 34).

Neal, B., Huang, C.-W., & Raghupathi, S. (2020). Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007* (cit. on pp. 5, 34).

Kuo, N. I.-H., Polizzotto, M. N., Finfer, S., Garcia, F., Sönnerborg, A., Zazzi, M., Böhm, M., Kaiser, R., Jorm, L., & Barbieri, S. (2022). The health gym: Synthetic health-related datasets for the development of reinforcement learning algorithms. *Scientific Data*, *9*(1), 693 (cit. on pp. 5, 33, 34).

Hernán, M. A. (2019). Comment: Spherical cows in a vacuum: Data analysis competitions for causal inference. *Statistical Science*, *34*(1), 69–71 (cit. on pp. 5, 34).

Bouneffouf, D., & Féraud, R. A tutorial on multi-armed bandit applications for large language models. In: In *Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining*. 2024, 6412–6413 (cit. on p. 7).

Bouneffouf, D., Rish, I., & Aggarwal, C. Survey on applications of multi-armed and contextual bandits. In: In *2020 ieee congress on evolutionary computation (cec)*. IEEE. 2020, 1–8 (cit. on p. 7).

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press. https://doi.org/10.1017/9781108571401 (cit. on p. 7).

Elena, G., Milos, K., & Eugene, I. (2021). Survey of multiarmed bandit algorithms applied to recommendation systems. *International Journal of Open Information Technologies*, *9*(4), 12–27 (cit. on p. 8).

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press. (Cit. on pp. 8, 10, 13).

Salomon, A., Audibert, J.-Y., & Alaoui, I. E. (2011). Regret lower bounds and extended upper confidence bounds policies in stochastic multi-armed bandit problem. *arXiv preprint arXiv:1112.3827* (cit. on p. 8).

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In: In *Proceedings of ieee 36th annual foundations of computer science*. IEEE. 1995, 322–331 (cit. on p. 8).

Bouneffouf, D. (2023). Multi-armed bandit problem and application (cit. on p. 9).

Lai, T. L., Robbins, H., et al. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, *6*(1), 4–22 (cit. on pp. 9, 17).

Agrawal, R. (1995). Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in applied probability*, *27*(4), 1054–1078 (cit. on p. 9).

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*, 235–256 (cit. on p. 9).

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*(Nov), 397–422 (cit. on pp. 9, 13).

Graepel, T., Candela, J. Q., Borchert, T., & Herbrich, R. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In: Omnipress. 2010 (cit. on p. 10).

Agrawal, S., & Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In: In *Conference on learning theory*. JMLR Workshop and Conference Proceedings. 2012, 39–1 (cit. on p. 10).

Kaufmann, E., Korda, N., & Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In: In *International conference on algorithmic learning theory*. Springer. 2012, 199–213 (cit. on p. 10).

Russo, D., & Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, *39*(4), 1221–1243 (cit. on p. 10).

Shang, X., Heide, R., Menard, P., Kaufmann, E., & Valko, M. Fixed-confidence guarantees for bayesian best-arm identification. In: In *International conference on artificial intelligence and statistics*. PMLR. 2020, 1823–1832 (cit. on p. 11).

Kalyanakrishnan, S., Tewari, A., Auer, P., & Stone, P. Pac subset selection in stochastic multi-armed bandits. In: In *Icml*. *12*. 2012, 655–662 (cit. on p. 12).

Gabillon, V., Ghavamzadeh, M., & Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in neural information processing systems*, *25* (cit. on p. 12).

Jamieson, K., & Nowak, R. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In: In *2014 48th annual conference on information sciences and systems (ciss)*. 2014, 1–6. https://doi.org/10.1109/CISS.2014.6814096 (cit. on p. 12).

Jun, K.-S., Jamieson, K., Nowak, R., & Zhu, X. Top arm identification in multi-armed bandits with batch arm pulls. In: In *Artificial intelligence and statistics*. PMLR. 2016, 139–148 (cit. on p. 12).

Jedra, Y., & Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, *33*, 10007–10017 (cit. on p. 12).

Langford, J., & Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, *20* (cit. on p. 13).

Ambinder, E. P. (2005). Electronic health records. *Journal of oncology practice*, *1*(2), 57 (cit. on pp. 15, 33).

Strehl, A., Langford, J., Li, L., & Kakade, S. M. (2010). Learning from logged implicit exploration data. *Advances in neural information processing systems*, *23* (cit. on p. 15).

Dudík, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* (cit. on p. 15).

Swaminathan, A., & Joachims, T. (2015a). Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, *16*(1), 1731–1755 (cit. on p. 15).

Swaminathan, A., & Joachims, T. (2015b). The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, *28* (cit. on p. 15).

Joachims, T., London, B., Su, Y., Swaminathan, A., & Wang, L. (2021). Recommendations as treatments. *AI Magazine*, *42*(3), 19–30 (cit. on p. 15).

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, *47*(260), 663–685 (cit. on p. 15).

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., & Zitouni, I. (2017). Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, *30* (cit. on p. 15).

Beygelzimer, A., & Langford, J. The offset tree for learning with partial labels. In: In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining*. 2009, 129–138 (cit. on p. 15).

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129 (cit. on p. 15).

Yin, M., & Wang, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, *34*, 4065–4078 (cit. on p. 15).

Oetomo, B., Perera, R. M., Borovica-Gajic, R., & Rubinstein, B. I. (2024). Warm-starting contextual bandits under latent reward scaling. *ICDM* (cit. on p. 16).

Huch, E. K., Shi, J., Abbott, M. R., Golbus, J. R., Moreno, A., & Dempsey, W. H. RoME: A robust mixed-effects bandit algorithm for optimizing mobile health interventions. In: In *The thirty-eighth annual conference on neural information processing systems*. 2024. https://openreview.net/forum?id=eKVugi5zr0 (cit. on p. 16).

Sen, R., Shanmugam, K., Kocaoglu, M., Dimakis, A., & Shakkottai, S. Contextual bandits with latent confounders: An nmf approach. In: In *Artificial intelligence and statistics*. PMLR. 2017, 518–527 (cit. on p. 16).

Kocák, T., Munos, R., Kveton, B., Agrawal, S., & Valko, M. (2020). Spectral bandits. *Journal of Machine Learning Research*, *21*(218), 1–44 (cit. on p. 16).

Kinyanjui, N. M., Carlsson, E., & Johansson, F. D. (2023). Fast treatment personalization with latent bandits in fixed-confidence pure exploration. *Transactions on Machine Learning Research* (cit. on p. 17).

Thomas M. Cover, J. A. T. (2005). Entropy, relative entropy, and mutual information. In *Elements of information theory* (pp. 13–55). John Wiley & Sons, Ltd. (Cit. on p. 17).

Kinyanjui, N. M., & Johansson, F. D. Adcb: An alzheimer's disease simulator for benchmarking observational estimators of causal effects. In: In *Conference on health, inference, and learning*. PMLR. 2022, 103–118 (cit. on pp. 19, 21, 34).

Russo, D. Simple bayesian algorithms for best arm identification. In: In *Conference on learning theory*. PMLR. 2016, 1417–1418 (cit. on p. 19).

Hyvarinen, A., & Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, *29* (cit. on p. 20).

Balcıoglu, A. Z., Mwai, N., Carlsson, E., & Johansson, F. D. (2025). Identifiable latent bandits: Leveraging observational data for personalized decision-making (cit. on p. 20).

Pearl, J. (2009). *Causality*. Cambridge university press. (Cit. on p. 20).

Mwai, N., Carlsson, E., & Johansson, F. D. (2025). Latent preference bandits (cit. on p. 22).

Barlow, R. E., & Brunk, H. D. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, *67*(337), 140–147 (cit. on p. 24).

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, *5*(4), 1–19 (cit. on p. 24).

Mwai, N., Malekipirbazari, M., & Johansson, F. D. (2025). Understanding exploration in bandits with switching constraints: A batched approach in fixed-confidence pure exploration (cit. on p. 28).

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, *3*(1), 1–9 (cit. on p. 33).

Yu, C., Liu, J., Nemati, S., & Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, *55*(1), 1–36 (cit. on p. 33).