Thesis for The Degree of Licentiate of Engineering

Improved sequential decision-making with structural priors

Enhanced treatment personalization with historical data

Newton Mwai

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2023

Improved sequential decision-making with structural priors
Enhanced treatment personalization with historical data

Newton Mwai

ISSN 1652-876X

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

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2023.

To my family.

Improved sequential decision-making with structural priors

Enhanced treatment personalization with historical data

Newton Mwai

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Personalizing treatments for patients involves a period where different treatments out of a set of available treatments are tried until an optimal treatment is found, for particular patient characteristics. To minimize suffering and other costs, it is critical to minimize this search. When treatments have primarily short-term effects, the search can be performed with multi-armed bandit algorithms (MABs). However, these typically require long exploration periods to guarantee optimality. With historical data, it is possible to recover a structure incorporating the prior knowledge of the types of patients that can be encountered, and the conditional reward models for those patient types. Such structural priors can be used to reduce the treatment exploration period for enhanced applicability in the real world. This thesis presents work on designing MAB algorithms that find optimal treatments quickly, by incorporating a structural prior for patient types in the form of a latent variable model. Theoretical guarantees for the algorithms, including a lower and a matching upper bound, and an empirical study is provided, showing that incorporating latent structural priors is beneficial. Another line of work in this thesis is the design of simulators for evaluating treatment policies and comparing algorithms. A new simulator for benchmarking estimators of causal effects, the Alzheimer's Disease Causal estimation Benchmark (ADCB) is presented. ADCB combines data-driven simulation with subject-matter knowledge for high realism and causal verifiability. The design of the simulator is discussed, and to demonstrate its utility, the results of a usage scenario for evaluating estimators of causal effects are outlined.

Keywords

Treatment personalization, fixed-confidence pure exploration, latent bandits, structural priors, historical data, policy optimization, benchmark simulators

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **Newton Mwai Kinyanjui**, 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.

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

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[**a**] **Newton Mwai Kinyanjui**, Fredrik D. Johansson, *ADCB: An Alzheimer's disease benchmark for evaluating observational estimators of causal effects Machine Learning for Health (ML4H) - Extended Abstract (December 2021), arXiv:2111.06811.*

Acknowledgment

I would like to express my deepest appreciation to my PhD advisor, Fredrik D. Johansson, for his helpful advice, unwavering guidance and helpful contributions in my research. I'm also extremely grateful to my co-advisor Morteza Chehreghani and my examiner Devdatt Dubhashi for their suggestions and invaluable insight into my PhD studies.

I am thankful for my PhD colleagues at DSAI who are pleasant to work with, among them Emil, Emilio, Hampus, Arman, Juan, Christopher, Riccardo, Markus, David, Firooz, Peter, Tobias, Simon, Mehrdad, Mena, Alexander, Filip, Daniel, Fazeleh, Hanna, Niklas and Hannes. Many thanks to the rest of the division including all faculty, post-docs and administrators. I am also grateful to have incredible office buddies Adam, Anton, Lena and Lovisa who lighten my days at the office. Special thanks to the fellow members of the WASP PhD council, Amandine, Anoud, Anton, Eduardo, Hooman, Mattias, Shuangshuang, and Kristin, who have been a delight to work with.

My PhD endeavour would not have been possible without my family and friends. I'm extremely grateful to my mum, Mumbi, for her thoughts and emotional support. I'd also like to thank my brother Kim and my sister Njoki for their relentless support. I very much appreciate my uncles Murimi, Muriithi, Jeremiah, Baru and their families for their profound belief in me. Special thanks to my cousins Shugu, Njoki, Leah, Mwai, Justin, Angela, Karen and Muthara for being my biggest cheer leaders. I'm also thankful to all my new friends in Gothenburg for their moral support, including Nader, Anna and Cate.

This work was supported in part by WASP (Wallenberg AI, Autonomous Systems and Software Program) funded by the Knut and Alice Wallenberg foundation. 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 to both organizations.

Contents

Part I

Summary

Chapter 1

Introduction

With recent advances in machine learning, there is growing interest in investigating how machine learning can be used in personalized medicine. Personalized medicine entails using individual patient charactersitics (e.g. a demographic variable, a biomarker, or a result of a diagnostic test) to identifiy the optimal treatment among a set of treatments at a point in time. Treatment personalization is particularly desirable in chronic diseases like Alzheimers Disease (AD), or Rheumatoid Arthritis (RA), where treatment and critical care is an ongoing process, typically undertaken over long periods of time [Chakraborty and Moodie, 2013].

An approach for exploring alternative treatments is multi-armed bandit algorithms (MABs) [Gittens and Dempster, 1979; Lai, Robbins et al., 1985]. MABs were originally motivated by medical applications in drug testing [Thompson, 1933] and they have a recent history of personalization applications, with popularity growing since the proposed application in personalized news recommendation [Li et al., 2010]. However, MABs tend to be sample-hungry (meaning that they usually have long exploration periods), to the point of being unsuitable for finding personalized treatments in real-world clinical settings. Because a long treatment search phase can prolong unnecessary suffering, it must be avoided and minimized whenever possible. Leveraging domain knowledge, for example prior knowledge of the types of patients that can be encountered is a possible solution to designing applicable MABs for treatment personalization.

Evaluating treatment policies and comparing treatment personalization algorithms is challenging, especially in the healthcare domain. Real-world implementation is often not an option and basing evaluation on observational data must rely on strong assumptions and access to large samples [Rosenbaum, Rosenbaum and Briskman, 2010]. As a result, methods researchers in these areas often turn to simulators for benchmarking [Dorie et al., 2019; Chan et al., 2021; Kuo et al., 2022]. However, benchmark simulators rarely incorporate causally verifiable subject-matter knowledge, which results in data that deficiently mirrors practice, particularly in healthcare [Hernán, 2019].

This licentiate thesis will discuss treatment personalization as sequential

decision-making in healthcare with historical data.

Here are the specific goals of this thesis:

G1: To introduce and discuss treatment personalization as an optimization objective in sequential decision-making in healthcare with multi-armed bandits (MABs).

G2: To introduce and discuss leveraging historical data in healthcare to design efficent treatment personalization MAB algorithms with structural priors.

G3: To discuss a method to build a semi-synthetic benchmark simulator that incorporates causally verifiable domain knowledge. This is to provide a realistic environment to compare sequential decision-making algorithms in healthcare.

Outline of Thesis: This thesis begins with a concise background introducing multi-armed bandit algorithms (MABs), and introducing how historical data is leveraged in designing personalization algorithms. A brief summary of the two appended papers included in this thesis is then presented. A concluding discussion on the current work, the limitations, and directions for future work then follows. The last section comprises the two appended papers.

Chapter 2

Background

2.1 Multi-armed bandits for treatment personalization

A multi-armed bandit problem is described as follows: An *agent* (a treatment personalization strategy) and an *environment* (patient) interact in sequence over T rounds. For each round, the agent takes an *action* $A_t \in \mathcal{A} = \{1, ..., K\}$ (e.g. trial treatment) and gets a *reward* $R_t \in \mathbb{R}$ (e.g. treatment outcomes). The aim of the agent is to take actions in a manner that maximizes the cumulative reward at the end of the T rounds.

More formally, for a set of K actions, $\mathcal{A} = \{1, ..., K\}$, the environment sets K reward probability distributions $\nu_1, ..., \nu_K$ with their respective parameters (which for Gaussian distributions with the same standard deviation σ are the means $\mu \in [0, 1]^K$; with $\mu(a) := \mathbb{E}[\nu_a]$ as the mean reward for action a). The problem then proceeds as shown below in Algorithm 1.

Algorithm 1: Multi-Armed Bandit Problem

for each round $t = 1, 2, ..., T$ **do**
- $a_t \in \mathcal{A}$ is chosen using an exploration-exploitation strategy
- a new independent, stochastic reward r_t is realized drawn from $\nu_{a_t, \hat{\mu}}$
- *updates are made for estimated parameters* $\hat{\mu}$

The key challenge in the MAB problem is what is referred to as the *exploration-exploitation dilemma*: deciding when to explore (choose new actions) or exploit (choose actions already selected). The challenge arises because the arm parameters $\hat{\mu}(a), a \in \mathcal{A}$ are unknown to the agent (algorithm used interchangeably) when the rounds start. The parameters $\hat{\mu}(a)$ have to be estimated over the rounds, and they are never perfect, even with a large number of rounds T [Elena, Milos and Eugene, 2021]. Therefore, a good MAB algorithm is an algorithm that has a provably good exploration-exploitation strategy, that optimize typically for the cumulative reward or regret. Popular MAB algorithms are derived variations of the Upper-confidence Bound (UCB) [Auer, Cesa-Bianchi and Fischer, 2002] and Thompson Sampling [Thompson, 1933;

Chapelle and Li, 2011; Agrawal and Goyal, 2012].

As mentioned, MAB algorithms are designed to maximize the cumulative reward, which is defined as:

$$S_T = \sum_{t=1}^{T} r_t$$

Regret: To measure how well the algorithm does well across different problem instances, a standard approach is to compare the MAB algorithm's cumulative reward to the *best-arm benchmark,* μ^* defined as the expected reward of always choosing the best action [Slivkins et al., 2019]. The regret at round T for a MAB algorithm is defined as:

$$R(T) = \mu^* T - \sum_{t=1}^{T} \mu(a_t)$$

where $\mu^* := \max_{a \in \mathcal{A}} \mu(a)$. Minimizing the regret is equivalent to maximizing the reward.

To give guarantees on how well a MAB (exploration-exploitation) strategy can perform, *regret bounds* are typically used. Regret lower bounds, which quant ify how challenging a bandit problem is in the worst-case sense (property of a policy, together with a set of environments and a horizon [Lattimore and Szepesvári, 2020]) are used together with upper bounds which describe how well the MAB strategy match the lower bound.

In regret minimization, there is the $\Omega(\sqrt{KT})$ lower bound; *For any bandit algorithm, there exists a problem instance such that* $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$. Variants of UCB and Thompson sampling strategies have been shown to have theoretical worst-case guarantees $\mathbb{E}[R(T)] \leq O(\sqrt{KT \log(T)})$, which match the lower bound up to logarithmic factors [Slivkins et al., 2019; Lattimore and Szepesvári, 2020].

2.1.1 Contextual bandits

In personalization applications, contextual multi-armed bandits (CMABs) have been introduced, which aim to answer not which action is best, but in what situation each action is best. CMABs follow when the MAB problem described above is augmented with *context*: information about each instance, for example individual patient characteristics. I.e, for a treatment personalization application, at the start of the round, the agent also observes a patient's context covariates (e.g., lab measurements) as a draw of a random variable $X \in \mathbb{R}^d$. The CMAB proceeds as outlined below in Algorithm 2.

Algorithm 2: Contextual Multi-Armed Bandit Problem

for each round $t = 1, 2, ..., T$ **do**
- algorithm observes a context x_t
- $a_t \in \mathcal{A}$ is chosen using an exploration-exploitation strategy
- a new independent, stochastic reward r_t is realized drawn from distribution (x_t, a_t)
- *updates are made for estimated parameters* $\hat{\mu}(a|x)$, $a \in \mathcal{A}, x \in \mathcal{X}$

The key challenge in contextual bandits is designing good reward models, that sufficiently estimate the underlying distribution. Several settings with specific assumptions on the reward models have been proposed for CMABs strategies in the regret minimization setting, among them under: Lipschitz assumption, linearity assumption, and others [Slivkins et al., 2019]. The LinUCB strategy [Abbasi-Yadkori, Pál and Szepesvári, 2011] as an example, is shown to have an expected regret upper bounded by $\mathbb{E}[R(T)] \leq Cd\sqrt{T}\log(TL)$, where $C > 0$ is a suitably large universal constant and L is a Lipschitz constant [Lattimore and Szepesvári, 2020].

Contextual bandits allow for complete personalization, where each instance (patient) is considered independent and the parameters are learned from scratch each time. However, this is naturally accompanied by long exploration periods, undesirable for application in realistic clinical settings, motivating the work in paper II, where a key difference is that the context considered is stationary.

2.1.2 Latent Bandits

The described contextual bandits setting assumes that all the contextual factors relevant for the problem are explicitly known, as well as their structures and values. However, it is plausible to think of scenarios when the rewards are influenced by unobserved factors, henceforth referred to as the *latent state*, $S \in \mathcal{S} = \{1, ..., M\}$. For example, in the treatment personalization case, the outcomes for a treatment could depend on an underlying, unobserved disease state. In this setting, the observed context provides some information to reduce the uncertainty of the underlying state. Under such a setting, Latent bandits [Maillard and Mannor, 2014; Zhou and Brunskill, 2016; Hong et al., 2020a; Hong et al., 2020b] have been proposed, where the goal of the agent is to identify the latent state, after which it can act optimally.

With historical data of previous agent-instance interactions, a latent state space and conditional reward models on the latent states can be constructed a priori. In our work, we refer to this as a latent variable model (LVM), which we assume is recovered perfectly. For new instances, this LVM can be used for personalization, and having it improves sample efficiency, hence shorter exploration.

More formally, in a latent bandit problem, there are $|S||K|$ probability distributions $\nu_{s,1}, ..., \nu_{s,K}$ with respective parameters $\mu_{s,1}, ..., \mu_{s,K}$ which are estimated a-priori (the LVM). A latent bandit problem proceeds as follows;

Algorithm 3: Latent Bandit Problem

for each round $t = 1, 2, ..., T$ **do**
- algorithm observes a context x_t
- algorithm estimates latent state s_t
- $a_t \in \mathcal{A}$ is chosen using an exploration-exploitation strategy
- a new independent, stochastic reward r_t is realized drawn from the distribution ν_{s_t,a_t}
- *updates are made for estimated latent state parameters* $\hat{\theta}$ *in* $p(s|h_t, \hat{\theta})$

Here, $H_t = (X_1, A_1, R_1, ..., X_t, A_t, R_t)$ denotes the history of context, actions and rewards, up to time t and h_t denotes the realisation of this random variable.

Hong et al. (2020a) provide algorithms with regret upper bounds $\mathbb{E}[R(T)] \leq O(\sqrt{MT\log(T)})$ which depend on the latent state dimension M, and they show that the bound can be tighter when $M \ll K$. It is of interest to the work presented in this thesis to see if these latent bandit ideas from the regret minimization setting could be applied in the pure exploration setting described below.

2.1.3 Pure exploration

In addition to the goal of maximizing the reward over the set of rounds, discussed in the previous sections, there is also a *pure exploration* MAB problem formulation where an agent aims to sample the available actions to gain relevant information about the environment quickly, regardless of the rewards [Kaufmann, 2020]. Two main settings in pure exploration are where an agent aims to identify the best arm either with the fewest rounds possible with a pre-specified probability of success (*fixed-confidence setting*), or where an agent aims to identify the best arm with the highest probability of success, with a fixed, pre-specified round horizon (*fixed-budget setting*). A concise elaboration on the fixed-confidence pure exploration follows, because this thesis comprises work in this setting.

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 and Kaufmann, 2016; 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:

$$\begin{aligned} &\underset{\phi}{\text{minimize}} && \mathbb{E}_\phi[\tau] \\ &\text{subject to} && P(\mu_{\hat{a}_\tau} < \mu^*) \leq \delta, \end{aligned} \tag{2.1}$$

For proposed fixed-confidence exploration strategies, theoretical guarantees are usually provided for the expected stopping time, $\mathbb{E}[\tau]$. Garivier and Kaufmann (2016) present a general lower bound for the expected stopping time as:

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

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

and where $d(.)$ is the KL-divergence. This implies an *asymptotic* ($\delta \to 0$) lower bound following from $\mathrm{kl}(\delta, 1-\delta) \sim \log(1/\delta)$:

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

It should be noted that optimal arm playing proportions, $w^*(\mu)$, for any strategy matching this bound can be computed by seeing the supremum above as a maximum, so $w^*(\mu) := \text{argmax}_{w \in \sum_K} \inf_{\lambda \in \text{Alt}(\mu)} (\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a))$. Also, the stopping time depends on the set of alternative states $\text{Alt}(\mu)$, which have optimal arms that differ from the optimal arm of μ, via the inverse relation of $T^*(\mu)$. Alternate states with close arm parameters induce a lower divergence, and therefore increased stopping time.

Garivier and Kaufmann (2016) also introduce the Track and Stop strategy with an asymptotic upper bound matching the lower bound above, which samples arms by playing arms in a manner that tracks the optimal proportions $w^*(\mu)$. Because their goal is to estimate arm parameters, it is different from the setting in paper II in this thesis where the optimal arm parameters are known and fixed, and where the optimal arm proportions $w^*_{x,a}(s)$ are from solving a different optimization problem. Also, the set of alternative states $\text{Alt}(\mu)$ is an infinite set in their setting, compared to a finite set $\text{Alt}_x(s)$ of size $\leq M - 1$ parameter vectors in ours. However, the track and stop strategy motivates one of the algorithms in paper II, the LLPT, and the accompanying asymptotic, $\delta \to 0$, sample complexity analysis.

2.2 Historical data in treatment personalization

There is extensive literature studying policy optimization with logged bandit feedback, also called off-policy learning or off-policy evaluation (OPE) [Strehl et al., 2010; Dudík, Langford and Li, 2011; Swaminathan and Joachims, 2015a, 2015b]. In this line of work, contextual agents are required to evaluate, improve and optimize a policy wholly in the offline dataset collected under a possibly unknown logging policy. The main challenge of evaluating policies offline is that the logging policy could be non-uniformly stochastic, which leads to challenges of bias in action selection, and variance in policy value estimation when there are small action propensities [Joachims et al., 2021].

Various methods have been proposed to solve the OPE challenges including *Inverse Propensity Score(IPS)* [Horvitz and Thompson, 1952; Swaminathan et al., 2017], *Direct Methods* [Beygelzimer and Langford, 2009], and *Doubly Robust* methods [Dudík, Langford and Li, 2011; Robins and Rotnitzky, 1995]. However, these methods are highly dependent on specific environments [Voloshin et al., 2019; Fu et al., 2021]. In addition, results with these methods require strong overlap assumptions in action selection of the logging policy, hence they are limited in realistic settings [Yin and Wang, 2021].

Because of the aforementioned challenges, we consider using learned *latent* structural priors from historical data that can be leveraged to minimize explorations in the online setting. This is in contrast to learning the policies purely offline in the historical data.

2.2.1 Historical data and environments for evaluating sequential decision-making in healthcare

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) that could be valuable if leveraged to design sequential decision-making systems to improve healthcare. However, due to challenges of accessibility attributable to 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 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 datasets [Dorie et al., 2019] which have many advantages but often lack the intricacies observed in reality [Hernán, 2019]. Other researchers have constructed purely data-driven benchmarks from actual samples, either simulating a subset or all of the observed variables using simulators fit to data [Chan et al., 2021; Kuo et al., 2022; Neal, Huang and Raghupathi, 2020]. However, purely data-driven approaches may fail to capture the causal structure of the systems they model. It is this thought that led to the work in paper I in this thesis, *ADCB: An Alzheimer's disease simulator for benchmarking observational estimators of causal effects.*

Chapter 3

Summary of Included Papers

3.1 ADCB: An Alzheimer's disease simulator for benchmarking observational estimators of causal effects

In this paper, described is a method for designing a semisynthetic benchmark simulator for longitudinal Alzheimer's disease data that incorporates verifiable causal domain knowledge. The goal for the project was 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.

Evaluating learned decision-making policies and observational estimators of causal effects is challenging, especially in the healthcare domain. Real-world implementation is often not an option and basing evaluation on observational data must rely on strong assumptions and access to large samples [Rosenbaum, Rosenbaum and Briskman, 2010]. As a result, methods researchers in these areas often turn to simulators for benchmarking [Dorie et al., 2019; Chan et al., 2021].

Simulated data have many advantages but often lack the intricacies observed in reality [Hernán, 2019]. For example, two of the most widely used benchmarks in the community studying causal effects, IHDP [Hill, 2011] and ACIC [Dorie et al., 2019], have response surfaces which are hand-crafted from simple mathematical building blocks. To improve on this, researchers have constructed benchmarks from actual samples, simulating a subset [Neal, Huang and Raghupathi, 2020] or all of the observed variables using simulators fit to data [Chan et al., 2021]. However, purely data-driven approaches may fail to capture the causal structure of the systems they model. [Hernán, 2019]

argued that, fundamentally, benchmarks must *"combine data analysis and subject-matter knowledge"*.

In this work, a simulator of clinical variables associated with Alzheimer's disease was designed, aimed to serve as a benchmark for causal effect estimation while modelling intricacies of healthcare data. The system was fit to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and ground hand-crafted components incorporating results from comparative treatment trials and observational treatment patterns. The simulator included parameters which alter the nature and difficulty of the causal inference tasks, such as latent variables, effect heterogeneity, length of observed subject history, behaviour policy and sample size. In addition to generating tunable high dimensional observational data with high realism based on a real world Alzheimer's setting, ADCB also generates longitudinal data that includes potential outcomes for all treatments at each step in the longitudinal axis.

The ADCB simulator was designed based on a longitudinal structural causal model between context, treatment and outcome variables. 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 the AD literature. This causal graph is shown in figure 3.1 below.

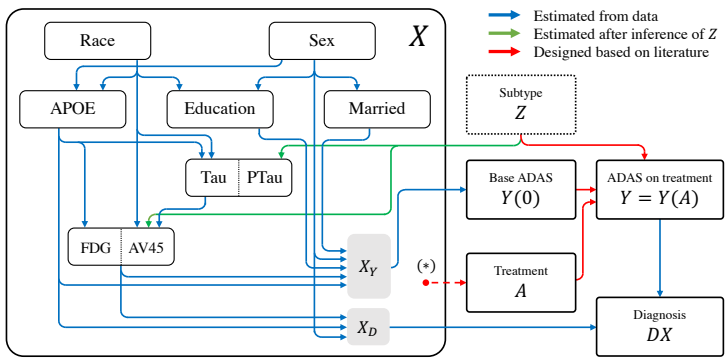
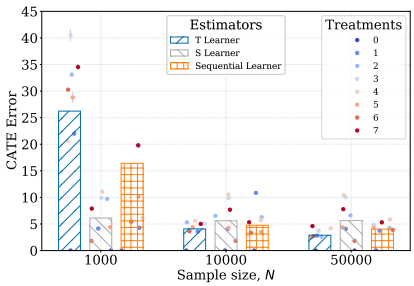


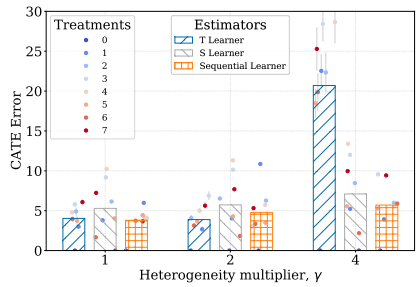
Figure 3.1: Assumed causal graph for the ADCB simulator at baseline. Arrows indicate causal dependencies, with colour representing how the mechanism was determined. Blue dependencies were completely estimated from data, green were fit once the subtype Z was inferred, and red were designed based on the Alzheimer's disease literature.

Using the ADCB simulator to compare standard estimators of causal effects, Conditional Average Treatment Effect (CATE) is then outlined, 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. This is illustrated in figure 3.2(a) and 3.2(b) below. Based on the results of these experiments, the benefits and limitations of our approach is discussed compared to existing simulators based on experimental data, hand-crafted mechanisms or learned functions.

At the time of publication, the benchmark simulator had only two latent



(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, $H = 3$

states. To make the system more realistic, the latent states were further expanded to six. In addition to this, a gym-like environment was built with logged data from the benchmark simulator, where MAB algorithms could be compared, and the environment was used for the experimental study in paper II.

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

This work outlines a method formulating treatment personalization as treatment search, solvable by fixed-confidence pure exploration. It also involves incorporation of a latent structural prior in the form of a Latent Variable Model (LVM) learnable from historical data to this setting, for fast treatment search. Two algorithms are designed and analysed, and empirical studies on a realistic AD environment with semisynthetic data from paper I are also given.

Personalizing treatments for patients often involves a period of trial-and-error search until an optimal choice is found. To minimize suffering and other costs, it is critical to make this process as short as possible. When treatments have primarily short-term effects, search can be performed with multi-armed bandits (MAB), but these typically require long exploration periods to guarantee optimality. In this work, MAB algorithms are designed, which provably identify optimal treatments quickly by leveraging prior knowledge of the types of decision processes (patients) we can encounter, in the form of a latent variable model. This is illustrated in figure 3.3 below.

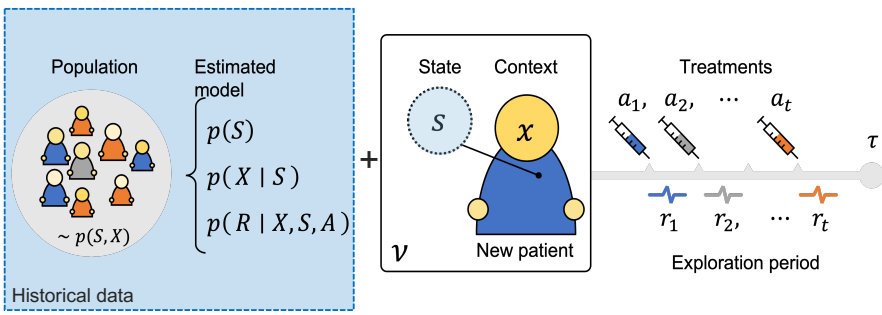
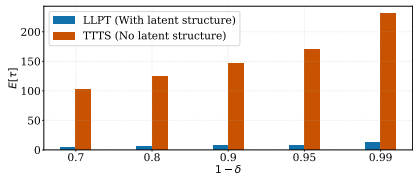


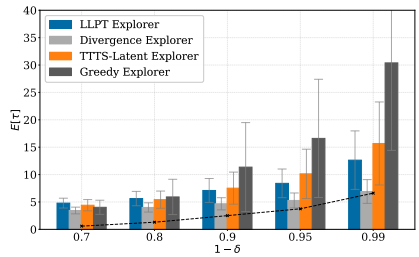
Figure 3.3: Illustration of the pure-exploration latent bandit problem and the example of treatment personalization. A population of patients have been observed in historical data to learn the distribution of latent states $P(S)$, $P(X|S)$ and the conditional reward the distribution $P(R|X,S,A)$. A new 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 design a MAB strategy ϕ to minimize the expected stopping time $\mathbb{E}[\tau]$ with a pre-specified confidence parameter $1 - \delta$ for new subjects with context X and unknown latent state S. In the healthcare example, this serves to minimize the search for optimal treatments, and thus minimize patient suffering in the treatment search phase while also ensuring that the algorithm commits to a good treatment after exploration, without treatment switches.

The main contributions are as follows: **1)** We propose a formulation of the personalized treatment search problem with known latent structure in the fixed-confidence pure-exploration setting. **2)** We prove a lower bound for the



(a) 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.



(b) Comparison of stopping time vs confidence $(1-\delta)$ for the algorithms. Our algorithms, LLPT Explorer and Divergence Explorer, have stopping times that are consistently lower.

search time of any algorithm in our latent bandit setting, and prove a matching upper bound for the Latent LP-based Track and Stop (LLPT) Explorer. **3)** We propose two algorithms, the LLPT Explorer and the Divergence Explorer. **4)** We perform an extensive empirical evaluation on a simulator of Alzheimer's disease and illustrate that our formulation and algorithms lead to a significantly reduced stopping time compared to classical pure-exploration algorithms in the MAB framework as illustrated in figure 3.4(a) and 3.4(b) above.

Chapter 4

Discussion and Future Work

In this thesis, leveraging structural priors, for example structural priors learned from historical data, in treatment personalization has been discussed. In addition, building benchmark simulators for sequential decision-making that incorporate subject-matter knowledge is included. The main research outcome from paper I was the Alzheimer's Disease Causal estimation Benchmark (ADCB), a simulator of clinical variables associated with Alzheimer's disease, aimed to serve as a benchmark for causal effect estimation, policy evaluation and algorithm comparison. The simulator is semi-synthetic: It has been designed by fitting a longitudinal causal model of patient variables to real data and also incorporating interventions with their average treatment effects, obtained from Alzheimer's Disease research literature. In paper II, the main research outcomes include the formulation of treatment as treatment search in the fixed-confidence pure exploration MAB setting, incorporation of a latent structural prior in the form of a Latent Variable Model (LVM) learnable from historical data, and two proposed MAB algorithms: the LLPT explorer and the Divergence explorer, of which the LLPT has been analyzed for its sample complexity. Results of an empirical study comparing the proposed algorithms to baselines in the ADCB environment has also been provided, showing that using the latent structure reduces the exploration period.

A limitation in the design of the simulator (paper I) and in the analysis in paper II is the simplifying assumption that the latent state is stationary. It is realistic to consider a dynamic latent state setting, where the latent state evolves over time with treatment. This could provide an interesting setting for design and analysis of algorithms in future work.

In paper II, our analysis is limited to the case in which the latent variable model is given and exact. When forced to estimate the model from historical data, sensitivity to misspecification or misestimation becomes a concern. Hong et al. (2020a) analysed latent bandits in regret minimization when the reward model is misspecified but the resulting bound suffers linear regret scaled by the error. A setting where a learner needs to recover the true model up to

some pre-specified precision is an interesting direction for future work. Another potential line of follow-up from paper II is analyzing the divergence explorer. Because the divergence explorer does not track optimal proportions like LLPT, we cannot rely on proof techniques from Track and Stop to analyze it, and its analysis is an interesting challenge for future work. This could also include an adaptation of the divergence explorer strategy to the regret minimization setting.

It would also be interesting to investigate other structural priors that can be incorporated in the treatment search MAB problem from historical data. An idea is to consider treatment monotonicity, for example regarding treatment monotone response structures and incorporating these in treatment search strategies.

In addition to studying the problem of treatment search in the best arm identification setting, it could be interesting to investigate treatment personalization with variations of contextual bandits in the regret minimization setting. Some interesting ideas to look into would be contextual bandits with missing data in the contexts or in the rewards. Another could be bandits with constraints reflecting real world scenarios, for example simulating domain guidelines, safety, or other costs in healthcare. It could also be interesting to investigate causality methods for off-policy learning towards enhanced contextual bandits that are realistic in real world clinical settings in healthcare.

Bibliography

Chakraborty, B., & Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*. Springer. (Cit. on p. 3).

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 p. 3).

Lai, T. L., Robbins, H., et al. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, *6*(1), 4–22 (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. 3, 5).

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 p. 3).

Rosenbaum, P. R., Rosenbaum, P., & Briskman. (2010). *Design of observational studies* (Vol. 10). Springer. (Cit. on pp. 3, 11).

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. 3, 10, 11).

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. 3, 10, 11).

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. 3, 10).

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. 3, 10, 11).

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. 5).

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*, 235–256 (cit. on p. 5).

Chapelle, O., & Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, *24* (cit. on p. 5).

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. 6).

Slivkins, A., et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, *12*(1-2), 1–286 (cit. on pp. 6, 7).

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press. (Cit. on pp. 6, 7).

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, *24* (cit. on p. 7).

Maillard, O.-A., & Mannor, S. (2014). Latent bandits. *31st International Conference on Machine Learning, ICML 2014* (cit. on p. 7).

Zhou, L., & Brunskill, E. (2016). Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743* (cit. on p. 7).

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. 7, 8, 17).

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 p. 7).

Kaufmann, E. (2020). *Contributions to the optimal solution of several bandit problems* (Doctoral dissertation). Université de Lille. (Cit. on p. 8).

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. 8, 9).

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. 8).

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. 9).

Dudík, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* (cit. on p. 9).

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. 9).

Swaminathan, A., & Joachims, T. (2015b). The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, *28* (cit. on p. 9).

Joachims, T., London, B., Su, Y., Swaminathan, A., & Wang, L. (2021). Recommendations as treatments. *AI Magazine*, *42*(3), 19–30 (cit. on p. 9).

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. 9).

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. 9).

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. 9).

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. 9).

Voloshin, C., Le, H. M., Jiang, N., & Yue, Y. (2019). Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854* (cit. on p. 9).

Fu, J., Norouzi, M., Nachum, O., Tucker, G., Wang, Z., Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., et al. (2021). Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596* (cit. on p. 9).

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. 9).

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. 10).

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. 10).

Neal, B., Huang, C.-W., & Raghupathi, S. (2020). Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007* (cit. on pp. 10, 11).

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240 (cit. on p. 11).

Part II

Appended Papers

Paper I

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

Newton Mwai Kinyanjui, and Fredrik D. Johansson

Proceedings of the Conference on Health, Inference, and Learning (April 2022), PMLR 174:103–118, 2022

Abstract

Simulators make unique benchmarks for causal effect estimation as they do not rely on unverifiable assumptions or the ability to intervene on real-world systems. This is especially important for estimators targeting healthcare applications as possibilities for experimentation are limited with good reason. We develop a simulator of clinical variables associated with Alzheimer's disease, aimed to serve as a benchmark for causal effect estimation while modeling intricacies of healthcare data. We fit the system to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and ground hand-crafted components in results from comparative treatment trials and observational treatment patterns. The simulator includes parameters which alter the nature and difficulty of the causal inference tasks, such as latent variables, effect heterogeneity, length of observed subject history, behavior policy and sample size. We use the simulator to compare standard estimators of average and conditional treatment effects.

Proceedings of Machine Learning Research 174:103–118, 2022 Conference on Health, Inference, and Learning (CHIL) 2022

ADCB: An Alzheimer's disease simulator for benchmarking observational estimators of causal effects

Newton Mwai Kinyanjui MWAI@CHALMERS.SE
Chalmers University of Technology, Sweden

Fredrik D. Johansson FREDRIK.JOHANSSON@CHALMERS.SE
Chalmers University of Technology, Sweden

Abstract

Simulators make unique benchmarks for causal effect estimation as they do not rely on unverifiable assumptions or the ability to intervene on real-world systems. This is especially important for estimators targeting healthcare applications as possibilities for experimentation are limited with good reason. We develop a simulator of clinical variables associated with Alzheimer's disease, aimed to serve as a benchmark for causal effect estimation while modeling intricacies of healthcare data. We fit the system to the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1] dataset and ground hand-crafted components in results from comparative treatment trials and observational treatment patterns. The simulator includes parameters which alter the nature and difficulty of the causal inference tasks, such as latent variables, effect heterogeneity, length of observed subject history, behavior policy and sample size. We use the simulator to compare standard estimators of average and conditional treatment effects.

Data and Code Availability We make use of publicly available longitudinal data, of both Alzheimer's disease (AD) patients and cognitively normal controls, from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). ADNI collects clinical data, neuroimaging data, genetic data, biological markers, and clinical and neuropsychological assessments from participants at different sites in the USA and Canada to study cognitive impairment and and AD. The cohorts used in this work were assembled from ADNI 1, 2, 3 and GO. We use trajectories of 870 unique patients, taking samples in 12-month intervals. An implementation of the simulator can be found at https://github.com/Healthy-AI/ADCB.

1. For the Alzheimer's Disease Neuroimaging Initiative: Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1. Introduction

Evaluating learned decision-making policies and observational estimators of causal effects is challenging, especially in the healthcare domain. Real-world implementation is often not an option and basing evaluation on observational data must rely on strong assumptions and access to large samples (Rosenbaum et al., 2010). As a result, methods researchers in these areas often turn to simulators for benchmarking (Dorie et al., 2019; Chan et al., 2021).

Simulated data have many advantages but often lack the intricacies observed in reality (Hernán, 2019). For example, two of the most widely used benchmarks in the community studying causal effects, IHDP (Hill, 2011) and ACIC (Dorie et al., 2019), have response surfaces which are hand-crafted from simple mathematical building blocks. To improve on this, researchers have constructed benchmarks from actual samples, simulating a subset (Neal et al., 2020) or all of the observed variables using simulators fit to data (Chan et al., 2021). However, purely data-driven approaches may fail to capture the causal structure of the systems they model. Hernán (2019) argued that, fundamentally, benchmarks must "combine data analysis and subject-matter knowledge".

We propose a new simulator for benchmarking estimators of causal effects, the Alzheimer's Disease Causal estimation Benchmark (ADCB). ADCB combines data-driven simulation with subject-matter

knowledge by fitting a longitudinal causal model of patient variables to real data: i) the simulator is based on a causal structure inferred by Alzheimer's disease experts, ii) average causal effects are based on published results from randomized controlled trials with heterogeneity introduced through an inferred latent variable, iii) overlap and variance in treatment choice is controlled by different behavior policies, and iv) the length of observed subject history is set by the user. This design provides users with tunable parameters which change properties of the system and the difficulty of the benchmark. We use the ADCB simulator to compare standard estimators of causal effects 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. Based on the results of these experiments, we discuss the benefits and limitations of our approach compared to existing simulators based on experimental data, hand-crafted mechanisms or learned functions.

2. Benchmarks for observational estimation of causal effects

Causal effect estimation studies the outcome $Y(a)$ of intervening with an action (treatment) $a \in \mathcal{A}$ (Rubin, 2005). Here, we define the causal effect of action a as

$$\Delta(a) \coloneqq Y(a) - Y(0),$$

the difference between the potential outcome of $A \leftarrow a$ and that of a baseline action $A \leftarrow 0$. In our setting, $\Delta(a)$ represents the benefit of using treatment a over no treatment. We consider k different actions from a discrete set $\mathcal{A} = \{0, ..., k-1\}$.

Due to the difficulty of trying out different treatments for the same subject under identical conditions, Δ itself is rarely identifiable. Instead, we represent the utility of actions using the *average treatment effect* (ATE), $\tau(a) = \mathbb{E}[\Delta(a)]$ and the *conditional average treatment effect* (CATE),

$$\tau(a \mid x) = \mathbb{E}[\Delta(a) \mid X = x],$$

in a context or stratum $x \in \mathcal{X}$. ATE and CATE measure how well action a performs on average in a population and in a stratum x, respectively. The context X may be a single vector-valued observation or a time-series representing patient history.

Observational estimation refers to estimating τ using samples (a, y, x) of actions, outcomes and context variables without controlling the actions. The following assumptions are sufficient for consistent, unbiased estimation in this setting (Rosenbaum et al., 2010).

Assumption 1 (Identifying assumptions) *Actions $A \in \mathcal{A}$, outcomes $Y \in \mathbb{R}$, a set of context variables X, and an adjustment set of variables $C \subseteq X$ are observed from a distribution $p(X, A, Y)$ such that the following conditions hold for all $a \in \mathcal{A}, c \in \mathcal{C}$,*

Consistency	$Y = Y(A)$
Exchangeability	$Y(a) \perp\!\!\!\perp A \mid C$
Overlap	$p(A = a \mid C = c) > 0$

A wealth of methods have been developed for estimating ATE and CATE under Assumption 1, see e.g., (Dorie et al., 2019; Künzel et al., 2019; Wager and Athey, 2018) for overviews. To assess the qualities of each estimator, various benchmark challenges have been developed (Dorie et al., 2019). See Section 6 for a more in-depth survey.

Fundamentally, the validity of Assumption 1 cannot be verified from data (Pearl, 2009), but must be argued from domain knowledge. Moreover, the assumptions guarantee *identification* of ATE and CATE, but not necessarily good *estimates* when sample sizes are small. Hence, observational data alone are insufficient to determine whether one estimate of a causal effect is more accurate than another. This motivates using simulators for benchmarking, where identifying assumptions can be satisfied by design.

A good benchmark allows users to identify strengths and weaknesses in estimators: Which estimators make efficient use of available data? How does performance scale with dimensionality or sample size? How sensitive are they to (partial) violations of identifying assumptions? Which results are robust to changes in causal structure? Answers to these questions will not be universal, they will depend on the application under study (Hernán, 2019). In this work, we target the healthcare domain, in the context of longitudinal data on clinical variables.

3. The ADCB simulator

Alzheimer's disease is the most common form of dementia, affecting tens of millions of people worldwide (Association, 2019). Despite its toll on public health and vast research investments over several decades, there is currently no cure for AD. Nevertheless, drugs that have disease-modifying effects, alleviating symptoms such as loss of cognitive function,

have shown promise in trials and in practice (Grossberg et al., 2019). For these reasons and more, AD makes an interesting setting for benchmarking causal effect estimators:

- AD is a progressive disorder, deteriorating the health of subjects over time. As a result, data is collected for the same subjects at several time points, allowing for comparing the performance of longitudinal models of causal effects.
- There is evidence that AD is composed of multiple disease subtypes. While the details remain unknown, disease subtypes provide a potential source of heterogeneity in patient outcomes.
- Current treatments are believed to be symptomatic—they affect only symptoms and not the underlying disease cause; their effects disappear once discontinued. This allows for easier attribution of effect to treatment.

The ADCB simulator is based on a longitudinal structural causal model between context, treatment and outcome variables. The remainder of the section describes the components of the simulator, starting with the patient covariates, the assumed causal graph, and a generalization to a longitudinal causal model. Framed boxes are used to indicate readily tunable parameters of the simulator.

3.1. Patient covariates X & outcomes Y

Subjects are represented by covariates $X \in \mathcal{R}^d$ consisting of demographics (sex, age, education level) and various genetic and biomarkers (Aβ plaques, Tau, APOE, FDG, AV45) whose detailed descriptions are provided in Appendix C. The specific variables used to model the time-varying context X_t in this work are presented in Figure 1. The severity of (suspected) Alzheimer's disease is primarily assessed based on cognitive function using tests such as the Alzheimer Disease Assessment Scale (ADAS) (Rosen et al., 1984). We use the ADAS13 variant as our base outcome at time t, $Y_t(0)$, as it has been found to better describe disease progression than the ADAS11 variant (Cho et al., 2021). ADAS13 scores take values between 0-85 where higher scores indicate worse cognitive function. ADNI also contains clinical diagnosis states $DX_t \in \{$Cognitively normal (CN), Mild cognitive impairment (MCI), Alzheimer's disease (AD)$\}$.

3.2. Disease subtype (latent state Z)

It is believed that there are multiple subtypes of Alzheimer's disease (Machado et al., 2020; Satone et al., 2018). One of the signs of this is that in subjects, the level of so-called Amyloid-β (Aβ) plaques form a clearly bimodal distribution, on the ratio of $(\frac{A\beta 42}{A\beta 40})$, see Figure 9 in Appendix. We posit that there are two types of subjects, as indicated by a binary variable $Z \in \{0, 1\}$, which, among other things, give rise to the two modes in the Aβ-ratio. To this end, we infer the subtype Z by fitting a Gaussian mixture model (GMM) with 2 components as in (Dansson et al., 2021) for the Aβ-ratio observations of patients at baseline. We assume that Z is stationary and use the value inferred by the GMM to label all observed trajectories. These values are then used to fit models of downstream variables.

3.3. Baseline Causal Graph

We start by positing a causal graph for the variables of interest at the baseline time point of observation, $t = 0$. A causal graph is a model of the (conditional) dependence structure of variables encoded in a directed acyclic graph, $\mathcal{G} = (\mathrm{V}, \mathrm{E})$ consisting of nodes V and edges, E (Koller and Friedman, 2009). The causal graph, illustrated in Figure 1, was inspired by the structure inferred from data in (Sood et al., 2020) and further verified by a clinically active domain expert in Alzheimer's disease. The graph represents causal relationships among random variables $R \in \{X(1), ..., X(d), A, Y, DX\}$, (where $X(j)$ is a covariate in the set X), each associated with a node in the graph, $V_R \in V$. An edge $(V_R, V_{R'}) \in E$ exists if R is a direct cause of R', and R is therefore a parent of R', $R \in Pa(R')$.

The mechanism for generating each variable is based either on models fit to the ADNI data, on handcrafted functions or on results from the AD literature. The graph is also presented as a table in the Appendix Table 5.

3.4. Longitudinal Model

The longitudinal model is formed by first repeating each variable, except the disease subtype Z, at each time step $t = 1, 2, ..., T$, maintaining the causal structure of the single-time graph in Figure 1. Then, each variable is connected to the previous instance of itself; e.g. Tau_t is assumed to be a direct cause of Tau_{t+1}, and so on. The parents of a variable X at time t is

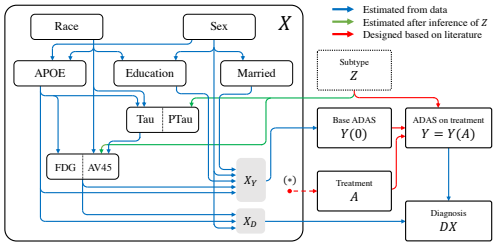


Figure 1: Assumed causal graph for a single time point at baseline. Arrows indicate causal dependencies, with color representing how the mechanism was determined. Blue dependencies were completely estimated from data, green were fit once the subtype Z was inferred, and red were designed based on the Alzheimer's disease literature.

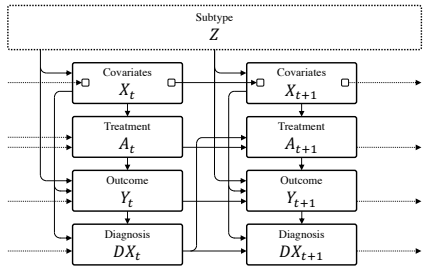


Figure 2: Temporal dependence between variables in the simulator. Each variable obeys the causal dependencies of Figure 1 in addition to depending on the previous value of itself. The small box in the set of covariates X indicates that each variable in the set depends only on the previous value of that specific variable. For example, Tau at time $t+1$ depends only on APOE and Race at time $t+1$, the subtype Z, and Tau at time t. Z is assumed stationary.

therefore the set defined as: $Pa(X_t) = Pa_t(X_0) \cup \{X_{t-1}\}$ where $Pa_t(X) = \{p_t : p_0 \in Pa(X_0)\}$. When used as a benchmark, the user may choose the causal effect of actions at any time point t as their target. The length $H \leq t$ of history used for estimation is a tunable parameter.

History Length, H. History of previous treatment records of a patient is valuable for causal effect estimators that incorporate history, because a longer horizon can increase the capacity to capture heterogeneity in causal effects.

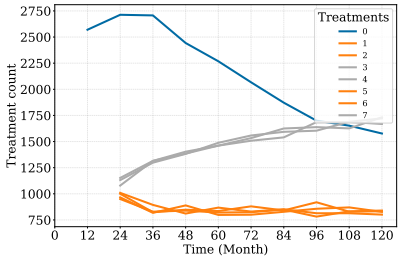
3.5. Treatment assignment A

ADNI does not include significant data on treatments and treatment response, which prevents direct data-driven design of the treatment assignment. Instead, we design policies for treatment assignment and treatment effects based on i) surveys of common treatments and ii) randomized controlled trials (RCT) of their effect. We begin with the former.

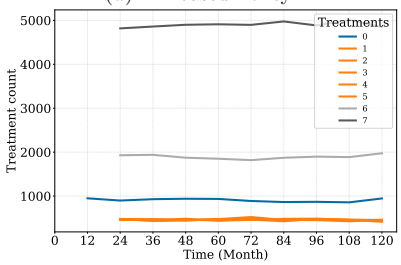
Existing AD drugs have been shown to have at least symptomatic cognitive effects (Livingston et al., 2017; Farlow et al., 2008). In this work, we model a range of such drugs $a = 1, ..., 7$, for which RCT results on treatment effects are available: Donepezil 5mg, Donepezil 10mg, Galantamine 24mg, Galantamine 32mg, Rivastigmine 12mg, Memantine 20mg, Memantine+ChEI, see (Grossberg et al., 2019) for an overview. We assume that the no-treatment option, $a = 0$, corresponds to observations in ADNI. We simulate treatments from two simple policies μ_B, described further below, whose characteristics are shown in Figure 3:

Diagnosis (DX)-based policy With this policy, treatments are assigned based on the diagnosis (DX) observed at the previous time step. We group treatments into 3 classes based on their treatment effect. Patients with mild diagnosis are assigned a randomly chosen treatment from the class with smallest ATE, those with moderate from the class with moderate ATE, and those with the most severe diagnosis from the class with the largest effect.

Hernandez Policy Having access to treatments in ADNI data would have enabled modeling of treatment propensities over the whole covariate set, deriving purely data-driven behavior policies using a much larger subset of covariates. In lieu of this, we draw from Hernandez et al. (2010) who similarly modeled the propensity of the treatments Cholinesterase inhibitors (ChEIs) and Memantine based on clinical variables with a multivariate logistic regression models, with ChEI or Memantine use as the outcome—



(a) DX-based Policy



(b) Hernandez policy

Figure 3: Treatment assignment characteristics of behavior policies over time. Same colour indicates treatments are considered to be in the same treatment assignment class.

we define a policy directly using the coefficients they learned. Treatments are grouped based on drug class $\in$ {ChEIs, Memantine, Combination therapy}. The learned policy depends also on cognitive scores MMSE and CDRSB, which are available in the ADNI database. We generate them in the same way as ADAS13.

Overlap strength ϵ The tuning parameter $\epsilon \in [0, 1]$ interpolates between a random policy ($\epsilon = 1$) and the policies above ($\epsilon = 0$) by assigning a random action with probability ϵ. Note that $\epsilon = 0$ does not always imply a lack of treatment group overlap, depending on the behavior policy.

3.6. Treatment effects Δ

Consistent with the AD literature, we assume that the effects of each existing drug a are primarily symptomatic and temporary, attenuating when treatment is stopped (Grossberg et al., 2019). In addition, we assume that the effect is stationary in time. To this end, we endow each treatment a with an additive effect $\Delta(a, Z)$, depending on the disease subtype Z, and posit that the cognitive function when on drug a is given by $Y_t(a) = \Delta(a, Z) + Y_t(0) + \epsilon_t$. This gives us the response surface on $Y_t = Y_t(A_t)$. $Y_t(0)$ is estimated from observations of the ADAS13 score and is simulated according to the causal graph. We discuss more general forms of treatment effects in Section 6.

To ground our model in domain knowledge, we design $\Delta(a, Z)$ such that the average effect $\tau(a) = \mathbb{E}[\Delta(a, Z)]$ is consistent with real-world effects on cognitive function (in the ADAS-Cog scale) estimated in RCTs (Grossberg et al., 2019). Recall that we define ATE relative to the no-treatment option. For a list of the estimated ATEs $\tau(a)$, for $a = 1, ..., k$, taken from the literature, see Appendix D.

Given the ATE $\tau(a)$ for a treatment a, heterogeneity is introduced through the subtype $z \in Z$. In this work, Z is binary, and we let each subtype-action pair (a, z) have HIGH or LOW effect, with multiplicative margin γ, such that the opposite subtype $(a, 1 - z)$ has HIGH effect, if (a, z) has LOW effect and vice versa.

$$\Delta(a, z) = \begin{cases} \frac{\tau(a)}{p(Z=z)+p(Z\neq z)\gamma}, & \text{if } \Delta(a, z) \text{ LOW} \\ \frac{\gamma\tau(a)}{p(Z=z)\gamma+p(Z\neq z)}, & \text{if } \Delta(a, z) \text{ HIGH} \end{cases}.$$

Whether $\Delta(a, z)$ is HIGH or LOW for a, z is determined by a look-up table that we designed which is presented in the Appendix, Table 3.

Treatment effect heterogeneity γ. The parameter $\gamma \geq 1$ controls heterogeneity in effect such that $\Delta(a, z) = \gamma\Delta(a, 1 - z)$ if $\Delta(a, z)$ is HIGH and vice versa. γ varies heterogeneity without changing the average treatment effect $\tau(a)$. $\gamma = 1$ results in no heterogeneity.

4. Fitting the simulator

Based on the causal the graph presented in Figure 1, we learn a joint distribution of the full set of set of observed variables $X, Y(0), DX$ by fitting each component of the Bayes factorization separately using a variable's parent set, $Pa(X_t)$. For each continuous (or discrete) attribute, a regression (or stochastic classification) model is fit with respect to its parents in the causal graph.

These models are first fit for the baseline timestep ($t = 0$) in patient trajectories for the purpose of i) generating the first time step further downstream in the generation process and ii) data imputation for missing values, as described in Appendix B. The marginal root nodes are sampled from a distribution inferred using the statistics observed in the data. Continuous covariates are further modeled with additive noise ζ sampled from a skewed normal distribution fit to the residuals of the regression, $r_i = y_i - f(x_i)$ where $f(x) \approx \mathbb{E}[Y|X = x]$ is learned from data.

The longitudinal model—the transition models for each variable—is fit similarly. For each covariate at time t, we assume that i) its value is dependent only on its parents in the causal graph at the time t as well as its previous value in the trajectory at time $t-1$. ii) the autoregression is stationary in time. A summary of the different models and their fit characteristics is described in Table 2.

For each time step, classifiers fit $P(X_t|Pa(X_t))$ and generation is done by sampling from this. The regressors fit $f(Pa(X_t)) = E[X_t|Pa(X_t)]$ and samples are generated by $f(Pa(X_t)) + \zeta$. With these models fit, hand-crafted components designed and tunable parameters $\{H, N, \gamma, \epsilon, T, \mu_B\}$ set, we generate N patient trajectories of T time steps with all variables $(Z, X_t, Y_t(0), A_t, Y_t(A_t), DX_t)$ through ancestral sampling.

4.1. ADNI and ADCB cohort statistics

Trajectories of 2254 subjects were downloaded from the ADNI database in December 2020. The full cohort was filtered for availability of measurements of Aβ40 and Aβ42 biomarkers at some point in their trajectory, leaving $n = 870$ subjects for fitting the simulator, 844 of which were observed at baseline. Overview statistics of these subjects at baseline are presented in Table 1. Trajectory lengths varied greatly among subjects, ranging from a single visit at baseline to a total of 8 visits (mean 1.7 visits). The longest trajectory length was 120 months (mean 14 months). Only subjects with observations for all simulator variables (except Z, A and $Y(a)$) were used for fitting baseline and autoregression covariate models. For longitudinal modeling, models were fit based on transitions between pairs of visits (0, 12), (12, 24), (24, 36), (36, 48) for observations present in both time points in the transition in the original data, which was a total of 127 samples.

Table 1: Cohort statistics for the first timestep ($t = 0$) for simulated (ADCB) and observed real-world subjects (ADNI). Continuous variables are described by mean (standard deviation) and categorical variables by count (frequency in %). Complete cohort statistics are provided in the Appendix table 4

	ADCB $t = 0$,	ADNI, $t = 0$
Demographics		
Gender		
Female	4807 (48.1%)	395 (46.8%)
Male	5193 (51.9%)	449 (53.2%)
Biomarkers		
Tau	286.0 (117.3)	279.6 (130.0)
PTau	27.9 (12.7)	26.7 (14.2)
FDG	1.3 (0.2)	1.2 (0.2)
AV45	1.2 (0.2)	1.2 (0.2)
APOE4		
0.0	4196 (42.0%)	460 (54.5%)
1.0	4460 (44.6%)	303 (35.9%)
2.0	1344 (13.4%)	81 (9.6%)
Outcomes		
ADAS13	16.4 (8.4)	15.4 (9.5)

4.2. Model fit for variables in causal graph

We evaluate the model fit on held-out data independently for each variable, as summarized in Table 2. The test split was always 20%. The overall predictability for baseline variables was low, with non-trivial accuracy attained only for a handful of the covariates, including diagnosis and AV45 levels. However, we remind the reader that accurate prediction is not the main goal of this step, but to learn a simulator with similar characteristics as the observed data. In Table 1 and Appendix Table 4 we show the first-order statistics for observed and generated data.

Autoregressors achieved significantly better results due to some variables being more or less static in time or varying very slowly. AV45 had surprisingly poor R^2 fit results for autoregression although the RMSE error was in the range of the standard deviation of the original data. The hyperparameters for the models were obtained by doing a grid search over combinations of parameters over Linear, Random Forest and Gradient Boosting estimators for each variable.

Table 2: Fit statistics for baseline and autoregression models on held-out data. Overall predictability was low at baseline and in autoregression for some continuous variables. This indicates that parents in the causal graph explain only a small amount of variance in the affected variables.. First-order statistics were well matched, see Table 1 and Appendix Table 4.

Target variable	Model	Baseline			Autoregression		
Classifiers		Acc	F1	# Classes	Acc	F1	# Classes
APOE4	KNN	45%	0.42	3	96%	0.94	3
Education (years)	Logistic Regression	21%	0.09	13	100%	1.00	10
Marital status	Logistic Regression	73%	0.62	5	96%	0.94	4
Diagnosis	Logistic Regression	63%	0.63	3	88%	0.87	3
Regressions		R^2	RMSE	σ_Y	R^2	RMSE	σ_Y
Tau	Random Forest	-1.13	105.35	133.43	0.73	47.81	117.95
PTau	Random Forest	-0.55	11.0	14	0.91	3.69	14
FDG	Gradient Boosting	-3.79	0.14	0.15	0.09	0.06	0.09
AV45	Random Forest	0.20	0.15	0.23	-82.03	0.12	0.12
ADAS13	Random Forest	0.21	6.36	9.6	0.55	4.09	6.3
CDRSB	Gradient Boosting	-0.06	1.26	1.5	-0.61	1.20	2.2
MMSE	Gradient Boosting	-0.56	2.03	2.6	-0.26	1.63	1.4

5. Using the benchmark

We run experiments aimed at exploring the utility of the simulator and its generated sequential trajectories in benchmarking causal effect estimators. The experiments compare estimators of the Conditional Average Treatment Effect (CATE) at a given timepoint $0 < t_s \leq T$. We run them in settings with decisions with single-time context X_{t_s} and in settings where context comprises a H-length history of context, treatment and outcome variables, and compare the mean-squared error in estimated CATE (also called precision of estimating heterogeneous effects (PEHE) (Hill, 2011)). Unless otherwise stated, Assumption 1 is satisfied in all experiments by giving estimators access to a valid adjustment set. The adjustment set includes all the covariates in the current demographics, the current biomarkers, the most recent outcome and most recent diagnosis for the DX-policy. A similar adjustment set is used for the Hernandez-based policy, without the most recent diagnosis and with CDRSB and MMSE scores, for validity.

The estimators presented are S-learners (treatment as a covariate) and T-learners (separate regression for each treatment) (Künzel et al., 2019) with Linear Regression, Gradient Boosting or Random Forest base learners, as well as a Sequential T-learner with an RNN base learner to enable incorporation of history. S- and T-learners are trained single-step and the sequential T-learner trained using a history sequence of time points $\{t = t_s - H, ..., t = t_s\}$.

We investigate the effects of sample size, overlap, heterogeneity, history length and confounding as outlined below. Results are from 10 repetitions in each configuration.

Sample size, N: Under Assumption 1, it is expected that the CATE estimation error shall decrease with higher sample sizes as the variance should decrease with more samples, until bias (model misspecification) dominates the error. Estimating CATE with different numbers of samples generated from ADCB is consistent with this across the estimators, as shown in Figure 4 where the base estimator for the T- and S-learners is a Gradient Boosting Regressor. The CATE error with 50,000 samples is comparable with the error using 10,000 samples, so the rest of the experiments have been run with 10,000 samples.

Overlap, ϵ: ADCB enables investigation of overlap with the tunable parameter ϵ, which varies the treatment assignment propensity characteristics of the treatment policies in Figure 3. As ϵ increases and selection bias decreases, the behavior policy approaches a uniform policy and it's expected that the CATE estimation error should decrease. This is ob-

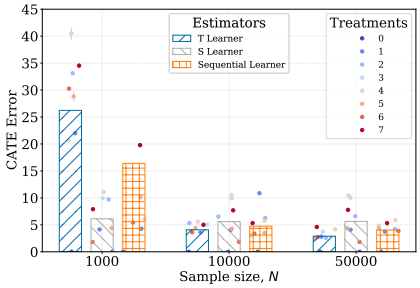
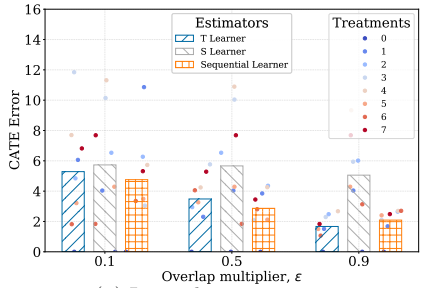
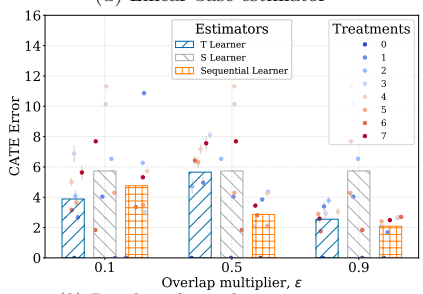


Figure 4: CATE mean squared error varying with sample size, N. ϵ=0.1, γ=2, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$



(a) Linear base estimator



(b) Random forest base estimator

Figure 5: Average (bars) and treatment-specific (dots) mean squared error in estimated CATE, varying with overlap multiplier, ϵ. γ=2, Sample size, $N = 10,000$, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$

served in the T-learner and the sequential learner (RNN), but the S-learner is constant through the three ϵ settings, as shown in Figure 5.

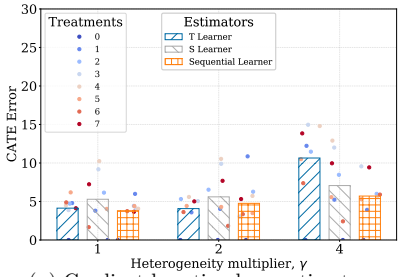
Heterogeneity, γ: With the tunable parameter γ, we can also vary the heterogeneity characteristics of the treatment policies. It is expected that the error should increase as the heterogeneity increases as higher heterogeneity may increase the variance, and the outcomes of actions become harder to predict. Our results in Figure 6 show this across two different base estimators (Gradient boosting and Random forest) in the T- and S- Learners.

History length, H: A key property of the ADCB simulator is access to history. Because physicians usually have access to historical records of a patient, they can use the historical records to personalize their treatment decisions. It is expected that using the history should decrease the error of the estimated CATE for the sequential learner that can incorporate history. This is because access to more history gives the estimator a higher chance of capturing heterogeneity. In Figure 7, the error for the T- and S- learners remains constant because they cannot make use of the history. The error is lowest with a history of length 2, possibly because the DX-based policy uses only the previous diagnosis. It would be interesting to investigate if other sequential estimators are better with longer histories.

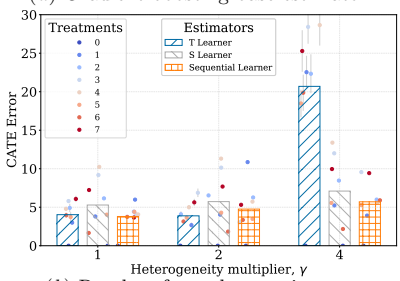
Confounding: Because we know the causal graph of the simulator, we can also investigate confounding effects, e.g. by adding current diagnosis in the adjustment set, which is a post-treatment collider variable, as shown in Figure 8. The estimators are affected differently by this confounding, with the sequential learner showing the highest error increase due to confounding. The T- and S- learners seem to be more robust with the T-learner being slightly more affected.

6. Discussion & Related work

The possibility of producing confounded evaluation metrics prevents using only observational data for benchmarking causal effect estimation, without relying on strong assumptions. There are two main approaches which do not rely on such assumptions: a) making use of data from randomized experiments, and b) simulating all or part the system under investigation, also called the Empirical Monte Carlo Study



(*a*) Gradient boosting base estimator



(*b*) Random forest base estimator

Figure 6: CATE error varying with heterogeneity, γ. ϵ=0.1, Sample size, $N = 10,000$, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$

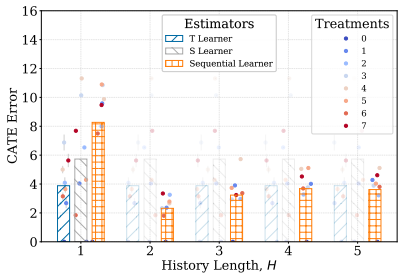


Figure 7: CATE error varying with sequence length, H. ϵ=0.1, γ=2, Sample size, $N = 10,000$, μ_B=DX-Based, $t_s = 5$. Estimators independent of history length are grayed out.

(EMCS) approach (Huber et al., 2013; Lechner and

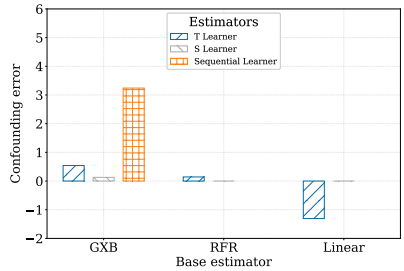


Figure 8: Excess error due to confounding, relative to CATE error of corresponding estimator, when post-treatment covariate DX is added to the adjustment set. ϵ=0.1, γ=2, Sample sizem, $N = 10,000$, μ_B=DX-Based, $t_s = 5$, History length, $H = 3$

Wunsch, 2013). See Gentzel et al. (2019) for a discussion of the pros and cons of each design.

Both methods have limitations that ADCB seeks to remedy. With randomized experiments data, as used in the Jobs dataset (Shalit et al., 2017) or by Neal et al. (2020), the data are guaranteed to be representative of the real world, but it is not possible to vary all characteristics of it, like the sample size or longitudinal horizon length. In contrast, for simulators, it is important to pay attention to the causal structure and mechanisms of the system which most often requires domain knowledge, without which high realism is not easily achievable.

If the goal of a benchmark is to evaluate individual-level or fine conditional treatment effects, access to counterfactual outcomes is required. The only way to reliably achieve this is to simulate the mechanism determining the outcome of interventions, which can be done in isolation or in addition to simulating the treatment assignment, as in the Causal Inference Benchmarking Framework by Shimoni et al. (2018), the Medkit-Learning environment (focused on reinforcement learning) (Chan et al., 2021), and in IHDP (Hill, 2011). Since the outcome mechanisms are often the main target of estimation, these simulations should be as realistic as possible for the domain they aim to represent. To this end, researchers have considered building their simulators on models fit to observational data (Neal et al., 2020; Chan et al., 2021). To incorporate domain knowl-

edge in simulating the outcomes and counterfactual outcomes, ADCB extends these approaches by using treatments and their corresponding effects from Alzheimer's literature, paired with causal generation of a common outcome measurement for cognitive function (ADAS13).

A drawback of simulated data is that, in many cases, simulators "tend to match the assumptions of the researcher" (Gentzel et al., 2019). This is especially problematic in cases where they are introduced to evaluate one particular estimator which may also match those assumptions. As a result, it is important that simulator-based benchmarks contain settings that tweak assumptions to appropriately test the robustness of estimators to these. ADCB enables settings with different configurations for overlap, sample size, patient heterogeneity, behaviour policy and longitudinal history length. Knowledge of the causal graph also enables investigation of estimator performance with confounding. It is also possible to violate consistency by introducing a probability that patients take the assigned treatment.

Limitations

Limitations of ADCB include the following. First, although the treatments, and treatment propensities in the case of the Hernandez policy (Hernandez et al., 2010), are obtained from literature, they are still simulated treatments not originally included in the ADNI data. As such, they may not reflect how subjects in the ADNI cohort would be treated under current practice. Further, behavior policies used only a single time-step context and not patients' entire history. Second, for the treatment effects, we use a simple bi-modal model of heterogeneity and the heterogeneity simulation assumes that heterogeneity is only due to latent covariates Z. A more expressive model would let heterogeneity depend also on X.

As pointed out earlier, several of the autoregressive models (for covariate transitions) had poor accuracy, in three cases with negative R^2. We believe that this could be improved in a future version of the simulator by changing the handling of missing data so that only the target variable for a particular edge in the causal graph is required observed when fitting the model. Currently, transition models are fit to complete cases.

For the presented usage scenario of comparing estimators, we only investigated a handful of simple estimators among a vast array of causal effect estimation methods. Finally, although the assumed causal graph was informed both by conversations with a domain practitioner and by data-driven estimates in (Sood et al., 2020), it would be of interest to test the sensitivity of treatment effect estimates to different adjustment sets or changes to the causal graph such as the addition of new links between covariate nodes.

7. Conclusion

We have introduced the Alzheimer's Disease Causal estimation Benchmark (ADCB), a simulator of clinical variables associated with Alzheimer's disease, aimed to serve as a benchmark for causal effect estimation and policy evaluation. The simulator is fit to covariates and outcomes from the ADNI database and uses models of treatments and treatment effects derived using subject-matter knowledge in the Alzheimer's disease literature. In addition to generating tunable high dimensional observational data with high realism based on a real world Alzheimer's setting, ADCB also generates longitudinal data that includes potential outcomes for all treatments at each step in the longitudinal axis. We also present a method to build semi-synthetic datasets by incorporating results from Alzheimer's literature which is highly effective in attaining realism, and encourages incorporation of inter-disciplinary domain-specific results in building synthetic datasets in machine learning and causal inference.

Usage scenarios for evaluating estimators of causal effects have been presented for varying configurations. Since ADCB generates longitudinal samples of all variables (patient covariates, treatments and outcomes) in the system, it can function as a generator of arbitrarily large observational (batch) data, as an online policy learning environment and for design and evaluation of causally adaptive treatment policies. More complex confounding models based on the AD literature will be explored in future iterations of the simulator, increasing the difficulty of the benchmark. To improve the predictability of the fitted models, the sample sizes will be increased in future iterations by expanding the filtering strategy for the samples included in the training sets.

Institutional Review Board (IRB)

All data collection by ADNI were approved by the Institutional Review Boards of all participating institutions. Written informed consent was obtained from

every research participant according to the Declaration of Helsinki and the Belmont Report.

Acknowledgments

This work was supported in part by WASP (Wallenberg AI, Autonomous Systems and Software Program) funded by the Knut and Alice Wallenberg foundation. 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.

Collection and sharing of the data used in this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Alzheimer's Association. 2019 alzheimer's disease facts and figures. *Alzheimer's & dementia*, 15(3): 321–387, 2019.

Alex J Chan, Ioana Bica, Alihan Huyuk, Daniel Jarrett, and Mihaela van der Schaar. The medkit-learn (ing) environment: Medical decision modelling through simulation. *arXiv preprint arXiv:2106.04240*, 2021.

Soo Hyun Cho, Sookyoung Woo, Changsoo Kim, Hee Jin Kim, Hyemin Jang, Byeong C Kim, Si Eun Kim, Seung Joo Kim, Jun Pyo Kim, Young Hee Jung, et al. Disease progression modelling from preclinical alzheimer's disease (ad) to ad dementia. *Scientific reports*, 11(1):1–10, 2021.

Hákon Valur Dansson, Lena Stempfle, Hildur Egilsdóttir, Alexander Schliep, Erik Portelius, Kaj Blennow, Henrik Zetterberg, and Fredrik D Johansson. Predicting progression & cognitive decline in amyloid-positive patients with alzheimer's disease. *Alzheimer's Research & Therapy*, 2021.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34 (1):43–68, 2019.

Martin R Farlow, Michael L Miller, and Vojislav Pejovic. Treatment options in alzheimer's disease: maximizing benefit, managing expectations. *Dementia and geriatric cognitive disorders*, 25(5): 408–422, 2008.

Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32:11722–11732, 2019.

George T Grossberg, Gary Tong, Anna D Burke, and Pierre N Tariot. Present algorithms and future treatments for alzheimer's disease. *Journal of Alzheimer's Disease*, 67(4):1157–1171, 2019.

Miguel A Hernán. Comment: Spherical cows in a vacuum: data analysis competitions for causal inference. *Statistical Science*, 34(1):69–71, 2019.

Santiago Hernandez, McKee J McClendon, Xiao-Hua Andrew Zhou, Michael Sachs, and Alan J

Lerner. Pharmacological treatment of alzheimer's disease: effect of race and demographic variables. *Journal of Alzheimer's Disease*, 19(2):665–672, 2010.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Martin Huber, Michael Lechner, and Conny Wunsch. The performance of estimators based on the propensity score. *Journal of Econometrics*, 175(1): 1–21, 2013.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165, 2019.

Michael Lechner and Conny Wunsch. Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics*, 21: 111–121, 2013.

Gill Livingston, A Sommerlad, V Orgeta, SG Costafreda, J Huntley, D Ames, C Ballard, S Banerjee, A Burns, J Cohen-Mansfield, et al. The lancet international commission on dementia prevention and care. *Lancet*, 390(10113): 2673–2734, 2017.

Alejandra Machado, Daniel Ferreira, Michel J Grothe, Helga Eyjolfsdottir, Per M Almqvist, Lena Cavallin, Göran Lind, Bengt Linderoth, Åke Seiger, Stefan Teipel, et al. The cholinergic system in subtypes of alzheimer's disease: an in vivo longitudinal mri study. *Alzheimer's research & therapy*, 12(1):1–11, 2020.

Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Stefan Klein, Daniel C Alexander, et al. Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease. *arXiv preprint arXiv:1805.03909*, 2018.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Wilma G Rosen, Richard C Mohs, and Kenneth L Davis. A new rating scale for alzheimer's disease. *The American journal of psychiatry*, 1984.

Paul R Rosenbaum, PR Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331, 2005.

Vipul Satone, Rachneet Kaur, Faraz Faghri, Mike A Nalls, Andrew B Singleton, and Roy H Campbell. Learning the progression and clinical subtypes of alzheimer's disease from longitudinal clinical data. *arXiv preprint arXiv:1812.00546*, 2018.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

Meemansa Sood, Akrishta Sahay, Reagon Karki, Mohammad Asif Emon, Henri Vrooman, Martin Hofmann-Apitius, and Holger Fröhlich. Realistic simulation of virtual multi-scale, multi-modal patient trajectories using bayesian networks and sparse auto-encoders. *Scientific reports*, 10(1):1–14, 2020.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(1): 1–67, 2011.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Appendix A. Empirical distribution of the Aβ ratio

The the ratio of $(\frac{A\beta-42}{A\beta-40})$ in subjects showing Amyloid-β (Aβ) plaques form a clearly bimodal distribution;

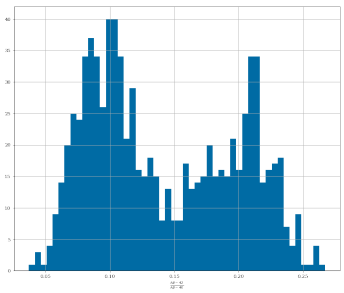


Figure 9: Empirical distribution of the Aβ ratio, used to infer latent disease subtype at baseline.

Appendix B. Imputation of missing data

The patient trajectories have significant missingness along the observation intervals. We impute the missing values based using a method inspired by Multivariate Imputation by Chained Equations(MICE) (Van Buuren and Groothuis-Oudshoorn, 2011), but the chaining is done with respect to a variable's parents in the causal graph. For each attribute with a missing value along the time trajectory, we use the model learned at baseline, from the causal graph, to impute the value for that particular attribute at a given timestep.

Appendix C. Patient covariates description

The subset of covariates used in this work includes the following and their descriptions as outlined in (Marinescu et al., 2018)

1. **FDG PET ROI averages:** Measure cell metabolism, where cells affected by AD show reduced metabolism
2. **AV45 PET ROI averages:** Measure amyloid-beta load in the brain, where amyloid-beta is a protein that mis-folds (i.e. its 3D structure is not properly constructed), which then leads to AD
3. **CSF biomarkers:** Amyloid and TAU levels in the cerebrospinal fluid (CSF)
4. Others:
   - **APOE status:** A gene that is a risk factor for developing AD
   - **Demographic information:** Gender, age, education, race, marital status
   - **Diagnosis:** Either Cognitively Normal (CN), Mild Cognitive Impairment (MCI) or Alzheimer's disease (AD).

Appendix D. Average Treatment Effects from Literature

Table 3: Average treatment effects (ATE), in terms of change in ADAS-Cog compared to no treatment, of various therapies from meta-analyses of clinical trials (Grossberg et al., 2019). Also shown is a look-up table for whether $\Delta(a, z)$ is HIGH or LOW

a	**Treatment**	**ATE** $\tau(a)$	$\Delta(a, z = 0)$
0	No treatment	0	-
1	Donepezil 5 mg	–1.95	L
2	Donepezil 10 mg	–2.48	L
3	Galantamine 24 mg	–3.03	H
4	Galantamine 32 mg	–3.20	H
5	Rivastigmine 12 mg	–2.01	L
6	Memantine 20 mg	–1.29	H
7	Memantine + ChEI	–2.64	L

Appendix E. Cohort statistics

Complete Cohort statistics for synthtetic and real-world cohorts are presented in Table 4.

Appendix F. Causal Graph

The expanded table for the causal graph in Figure 1 is presented in Table 5.

Table 4: Cohort statistics for the first timestep (T=1) for simulated (ADCB) and observed real-world subjects (ADNI). Continuous variables are described by mean (standard deviation) and categorical variables by count (frequency in %).

	ADCB T=1, n=10000	ADNI T=1, n=844
Demographics		
Gender		
Female	4807 (48.1%)	395 (46.8%)
Male	5193 (51.9%)	449 (53.2%)
Marital status		
Divorced	7572 (75.7%)	634 (75.1%)
Married	387 (3.9%)	29 (3.4%)
Never married	1098 (11.0%)	96 (11.4%)
Unknown	889 (8.9%)	80 (9.5%)
Widowed	54 (0.5%)	5 (0.6%)
Ethnicity		
Hisp/Latino	341 (3.4%)	30 (3.6%)
Not Hisp/Latino	9605 (96.0%)	809 (95.9%)
Unknown	54 (0.5%)	5 (0.6%)
Race		
Am Indian/Alaskan	9269 (92.7%)	783 (92.8%)
Asian	384 (3.8%)	31 (3.7%)
Black	148 (1.5%)	13 (1.5%)
Hawaiian/Other PI	17 (0.2%)	1 (0.1%)
More than one	137 (1.4%)	12 (1.4%)
Unknown	18 (0.2%)	2 (0.2%)
White	27 (0.3%)	2 (0.2%)
Education	13.2 (2.7)	13.3 (2.6)
Biomarkers		
Tau	286.0 (117.3)	279.6 (130.0)
PTau	27.9 (12.7)	26.7 (14.2)
FDG	1.3 (0.2)	1.2 (0.2)
AV45	1.2 (0.2)	1.2 (0.2)
APOE4		
0.0	4196 (42.0%)	460 (54.5%)
1.0	4460 (44.6%)	303 (35.9%)
2.0	1344 (13.4%)	81 (9.6%)
Outcomes		
ADAS13	16.4 (8.4)	15.4 (9.5)
MMSE	27.5 (2.0)	27.6 (2.5)
CDRSB	2.0 (1.3)	1.5 (1.7)
Diagnosis		
CN	2700 (27.0%)	275 (32.6%)
Dementia	5817 (58.2%)	438 (51.9%)
MCI	1483 (14.8%)	131 (15.5%)
Subtype, Z		
Subtype, Z	4282 (42.8%)	- (-)

Table 5: Expanded table for the causal graph in Figure 1 at baseline (t=0). For each time step, classifiers fit $P(X_t|Pa(X_t))$ and generation is done by sampling from this. The regressors fit $f(Pa(X_t)) = E[X_t|Pa(X_t)]$ and samples are generated by $f(Pa(X_t)) + \epsilon$. The models are the best performers after a grid search over hyperparameters.

Target variable (X)	Model	Direct causes at baseline ($Pa(X_{t=0})$)
Classifiers		
APOE4	KNN	Ethnicity, Race, Gender
Education (years)	Logistic Regression	Ethnicity, Race, Gender
Marital status	Logistic Regression	Gender
Diagnosis	Logistic Regression	Ethnicity, Race, Gender, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13
Regressions		
Tau	Random Forest	Ethnicity, Race, Gender, Z, APOE4
PTau	Random Forest	Ethnicity, Race, Gender, Z, APOE4
FDG	Gradient Boosting	Ethnicity, Race, Z, Tau, PTau, APOE4
AV45	Random Forest	Ethnicity, Race, Z, Tau, PTau, APOE4
ADAS13	Random Forest	Ethnicity, Race, Education, Gender, Marital status, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13
CDRSB	Gradient Boosting	Ethnicity, Race, Education, Gender, Marital status, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13
MMSE	Gradient Boosting	Ethnicity, Race, Education, Gender, Marital status, Z, Tau, PTau, APOE4, FDG, AV45, ADAS13

Paper II

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

Newton Mwai Kinyanjui, Emil Carlsson, Fredrik D. Johansson

Transactions on Machine Learning Research Journal. April 2023

Abstract

Personalizing treatments for patients often involves a period of trial-and-error search until an optimal choice is found. To minimize suffering and other costs, it is critical to make this process as short as possible. When treatments have primarily short-term effects, search can be performed with multi-armed bandits (MAB), but these typically require long exploration periods to guarantee optimality. In this work, we design MAB algorithms which provably identify optimal treatments quickly by leveraging prior knowledge of the types of decision processes (patients) we can encounter, in the form of a latent variable model. We present two algorithms, the Latent LP-based Track and Stop (LLPT) explorer and the Divergence Explorer for this setting: fixed-confidence pure-exploration latent bandits. We give a lower bound on the stopping time of any algorithm which is correct at a given certainty level, and prove that the expected stopping time of the LLPT Explorer matches the lower bound in the high-certainty limit. Finally, we present results from an experimental study based on realistic simulation data for Alzheimer's disease, demonstrating that our formulation and algorithms lead to a significantly reduced stopping time.

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

Newton Mwai *mwai@chalmers.se*
Department of Computer Science and Engineering
Chalmers University of Technology

Emil Carlsson *caremil@chalmers.se*
Department of Computer Science and Engineering
Chalmers University of Technology

Fredrik D. Johansson *fredrik.johansson@chalmers.se*
Department of Computer Science and Engineering
Chalmers University of Technology

Reviewed on OpenReview: *https://openreview.net/forum?id=NNRIGE8bvF*

Abstract

Personalizing treatments for patients often involves a period of trial-and-error search until an optimal choice is found. To minimize suffering and other costs, it is critical to make this process as short as possible. When treatments have primarily short-term effects, search can be performed with multi-armed bandits (MAB), but these typically require long exploration periods to guarantee optimality. In this work, we design MAB algorithms which provably identify optimal treatments quickly by leveraging prior knowledge of the types of decision processes (patients) we can encounter, in the form of a latent variable model. We present two algorithms, the Latent LP-based Track and Stop (LLPT) explorer and the Divergence Explorer for this setting: fixed-confidence pure-exploration latent bandits. We give a lower bound on the stopping time of any algorithm which is correct at a given certainty level, and prove that the expected stopping time of the LLPT Explorer matches the lower bound in the high-certainty limit. Finally, we present results from an experimental study based on realistic simulation data for Alzheimer's disease, demonstrating that our formulation and algorithms lead to a significantly reduced stopping time.

1 Introduction

There is growing interest in using machine learning for personalizing medical treatments to account for heterogeneity in patients' responses. Finding a suitable choice for an individual often involves a phase of trial and error before settling on a therapy that works for them, especially in the treatment of chronic diseases (Fraenkel et al., 2021; Stern, 2009). In rheumatoid arthritis, for example, when first and second-line treatment fails, there is large variability in the choice of next therapy, and several drugs may be considered equally good choices a priori (Zink et al., 2001). Further, switching therapies has associated costs: every time a therapy is changed, the patient has to get used to the new therapy and its potential side effects. It is therefore desirable to minimize such switches, even if changes are to other equally good treatments after a treatment has been identified in the search phase. Learning algorithms could improve the efficiency of this search, reducing the number of avoidable trials (Chakraborty and Moodie, 2013).

A classical framework for exploring alternative treatments is Multi-armed Bandits (MAB) (Gittens and Dempster, 1979; Lai and Robbins, 1985), originally motivated by reducing suffering in drug testing (Thompson, 1933). However, MABs tend to be sample-hungry to the point of being unsuitable for finding personalized treatments in real-world clinical settings. Because a long search phase can prolong unnecessary suffering, it

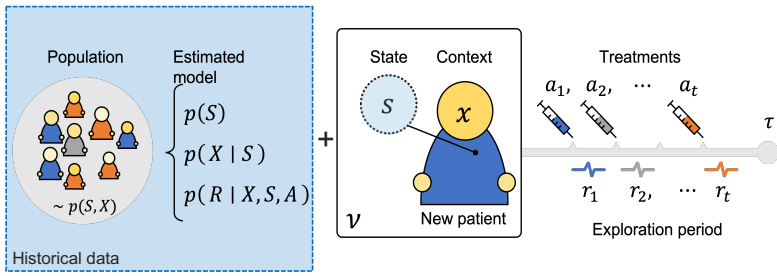


Figure 1: Illustration of the pure-exploration latent bandit problem and the example of treatment personalization. A population of patients have been observed in historical data to learn the distribution of latent states $P(S)$, $P(X|S)$ and the conditional reward the distribution $P(R|X, S, A)$. A new patient, represented by the instance $\nu = (x, s)$ is treated with actions a_t, observing rewards r_t until the stopping time τ.

must be avoided and minimized whenever possible. Existing methods for the fixed-confidence pure exploration setting in MABs, which aim to minimize the time it takes to find an optimal treatment at a given certainty level (Even-Dar et al., 2006; Garivier and Kaufmann, 2016; Russo, 2016; Shang et al., 2020) also yield long exploration phases.

One reason for the long exploration of bandit algorithms is that each instance—each patient, in our example—is treated as independent, learning parameters from scratch each time. This allows for complete personalization, often incorporating contextual or side information (Li et al., 2010; Chu et al., 2011), but disregards any similarities between instances. For many conditions, differences in responses (rewards) to treatment between patients are believed to be explained by a small number of disease subtypes (Devi and Scheltens, 2018; Borish and Culp, 2008). Thus, for a patient with a known subtype, an optimal treatment could be identified from the treatment responses of previous patients with the same subtype.

The subtype of a patient may be viewed as a latent state, as it is unobserved at the start of treatment, but manifests in a patient's responses to different therapies. With access to data on the treatment of previous patients, it is possible to fit a model of the distribution of latent states and their association with actions and rewards, for instance with variational inference methods (Kingma and Welling, 2013; Jang et al., 2016). Given such a model, for a new patient (bandit instance), our task becomes to identify which latent state they belong to, see Figure 1. Latent Bandits and recent iterations formalize this idea but are limited to regret minimization, aiming to minimize the regret compared to optimal actions over a possibly infinite period (Maillard and Mannor, 2014; Zhou and Brunskill, 2016; Hong et al., 2020a;b; Kwon et al., 2021). This differs from our goal of finding the optimal treatments within a desirably short exploration period, while also ensuring that the algorithm commits to a good treatment after exploration, without treatment switches.

In this work, we derive fixed-confidence pure-exploration bandit algorithms which aim to minimize the number of trials required to find an individual-optimal treatment by incorporating existing knowledge of latent structure.

Main contributions. 1) We propose a formulation of the personalized treatment search problem with known latent structure in the fixed-confidence pure-exploration setting (Section 2). **2)** We prove a lower bound for the search time of any algorithm in our latent bandit setting and prove a matching upper bound for the Latent LP-based Track and Stop (LLPT) Explorer (Section 3, 5). **3)** We present two algorithms, the LLPT Explorer and the Divergence Explorer (Section 4). **4)** We perform an extensive empirical evaluation on a simulator of Alzheimer's disease and illustrate that our formulation and algorithms lead to a significantly reduced stopping time compared to classical pure-exploration algorithms in the MAB framework (Section 6).

2 Problem formulation

We think of a treatment personalization strategy as an agent which interacts with a patient over $t \in \mathbb{N}$ rounds, aiming to try as few treatments as possible before the best possible treatment has been identified with a confidence level of at least $1 - \delta$, for a pre-specified $\delta > 0$. At the start of the sequence, the agent observes a patient's context covariates (e.g., lab measurements) as a draw of a random variable $X^1 \in \mathbb{R}^d$. Then, at each step $t = 1, 2, ...$, the agent takes an action $A_t \in \mathcal{A} = \{1, ..., K\}$ (trial treatment) and gets a reward $R_t \in \mathbb{R}$ (treatment outcomes). When an optimal treatment has been found, exploration stops and the agent recommends this treatment. The setting is illustrated in Figure 1. In the multi-armed bandit literature this setting is called fixed-confidence pure exploration (Garivier and Kaufmann, 2016; Shang et al., 2020).

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. Our goal is to design a strategy ϕ to minimize the expected stopping time $\mathbb{E}[\tau]$. In our healthcare example, this serves to minimize the search for optimal treatments, and thus minimize patient suffering in the treatment search phase while also ensuring that the algorithm commits to a good treatment after exploration, without treatment switches.

Even for state-of-the-art pure exploration algorithms, the necessary exploration tends to be long in realistic settings (see Figure 2). To overcome this, we will make structural assumptions about contexts, actions and rewards regarding patient similarity. In our healthcare example, it is plausible that a new patient (bandit problem instance) shares significant similarity with historical patients (logged bandit data), and that the optimal treatment for them is the same as for similar patients. However, in many domains, the context X is not sufficient to identify optimal treatment since it does not account for all individual variation (Håkansson et al., 2020). To account for remaining individual variation between patients with the same X, we will assume that there is a finite number of latent states $S \in \mathcal{S} = \{1, ..., M\}$, e.g., patient types, which cannot be directly observed. Thus, the optimal treatment is determined by the context X and the latent state S: two instances (e.g., two patients) are similar if they have the same context and latent state (e.g., disease subtype).

Identifying the true latent state S is sufficient but not strictly necessary to solve our problem. For successful treatment, we are only interested to identify the optimal treatment at exploration stop, $\hat{a}_\tau$. Therefore, it is not necessary to estimate the correct latent state, but the set of latent states that have the same optimal arm. Having context X is desirable as it helps reduce the number of trials if it is informative of the underlying latent state S, with unexplained variation further discoverable by trying different treatments.

A latent variable model (LVM) of the distribution of latent states S, contexts X, actions A and rewards R can be estimated from historical data and used to speed up exploration for a new subject. Maillard and Mannor (2014) and Hong et al. (2020a) made use of LVMs for "Latent Bandits" in the related setting of regret minimization. As these algorithms do not come with stopping and/or recommendation rules, they are not applicable to the fixed-confidence setting where the goal is to terminate search as quickly as possible.

In the MAB formalism, our problem can be defined as fixed-confidence pure-exploration latent bandits with a single initial context. In doing so, we assume that the latent subtype and the distributions of rewards is unaffected by time and previous actions. This is plausible for conditions treated with symptomatic therapies, such as for chronic degenerative disease like AD or Rheumatoid Arthritis (RA), where treatments typically target the symptoms and not the underlying disease pathology (Fish et al., 2019). Under these assumptions, the optimal choice of treatment remains fixed through exploration.

2.1 Fixed-confidence pure-exploration latent bandits

Given a state s, a context x, and an action a, let

$$\mu_{a,x,s} := \mathbb{E}[R \mid A = a, X = x, S = s]$$

[1]By convention, we use capital letters for random variables and lowercase for observed variables

denote the expected reward for that action, and let

$$\mu^*_{x,s} = \max_a \mu_{a,x,s} \quad \text{and} \quad a^*_{x,s} = \arg\max_a \mu_{a,x,s}$$

denote, respectively, the optimal expected reward and arm in latent state s and observed context x. We assume that the maximizer $a^*_{x,s}$ is a single action for each state-context pair (x, s), but our arguments can be generalized to the case with multiple optimal actions. Further, let $H_t = (X, A_1, R_1, ..., A_t, R_t)$ denote the history of context, actions and rewards, up to time t, letting $H_0 = (X)$. The utility of the context X is in computation of the likelihood $P(s|H_t)$ and this is agnostic of either finite or infinite context assuming that a good model of the likelihood is known.

Our goal is to design a search strategy ϕ *to minimize the expected number of trials* τ *required to identify an optimal action, with confidence at least* $1 - \delta$*, for new subjects with context* X *and unknown latent state* S*.*

$$\begin{aligned} &\underset{\phi}{\text{minimize}} && \mathbb{E}_{\phi,S,H_\tau}[\tau] \\ &\text{subject to} && P(\mu_{\hat{a}_\tau,x,s} < \mu^*_{x,s} \mid X = x, S = s) \leq \delta, \ \ \forall x, s \end{aligned} \tag{1}$$

We say that a search strategy ϕ is δ-PAC if the error probability is bounded by δ. Here, this is captured by our constraint, $\forall x, s : P(\mu_{\hat{a}_\tau,x,s} < \mu^*_{x,s} \mid X = x, S = s) \leq \delta$, as long as the probability model is correct.

In equation 1, we minimize the expected stopping time (e.g., over a population of patients) while satisfying instance-dependent constraints (per patient). We justify this formalization by noting that, in our running example, any single patient will have a single random stopping time, which we can estimate and analyze only in expectation. However, it is desirable and possible to guarantee, per patient, that our confidence exceeds $1 - \delta$ whenever we stop.

We assume that a model $\mathcal{M}_\theta = \{p_\theta(S), p_\theta(X \mid S), p_\theta(R \mid A, X, S)\}$ of the marginal state probability $p(S)$ and the likelihood of observed variables under S, including the set of reward means $\mu_{a,x,s}$, is *available when search begins*, akin to Hong et al. (2020a). This means that once s is known, so is the optimal arm in s, and no further exploration is necessary. Such a model can be learned from logged bandit instances, for example, using a variational autoencoder (Kingma and Welling, 2013), but this is outside the scope of this work.

For simplicity, we will assume that all reward distributions are stationary in time and Gaussian with equal variance σ^2, that is, given $A_t = a, X = x, S = s$, for all t

$$R_t \sim \mathcal{N}(\mu_{a,x,s}, \sigma^2) \ .$$

The algorithms presented in Section 4 are applicable in the non-Gaussian case as well, assuming that the reward distribution is known through $\mathcal{M}_\theta$, but our analysis in Section 5 is limited to Gaussian rewards for now. Our analysis makes heavy use of the Kullback-Leibler (KL) divergence, and we will adopt the notation $\mathrm{KL}(\mu_{a,x,s} \,\|\, \mu_{a,x,s'}) = \mathrm{KL}(p(R \mid a, x, s) \,\|\, p(R \mid a, x, s'))$ for the KL-divergence between the two Gaussian rewards for arm a under states s, s' with equal variance σ^2 and means as indicated.

3 Lower bound on stopping time

To serve as benchmark for our algorithms, we derive a lower bound on the worst-case solution to objective equation 1 for any algorithm which satisfies its constraints.

The seminal work of Kaufmann et al. (2016) presented a general inequality from which one can derive lower bounds for δ-PAC algorithms in the best-arm identification framework. In lemma 1, we present a variant of their key result, adapted to our latent bandit setting. For brevity, we let

$$\rho(x; s, s') = \log[p(x \mid s)/p(x \mid s')]$$

denote the log-likelihood ratio of the observed context x under latent states s and s', and use the shorthand

$$\mathbb{KL}^{R,a,x}_{s,s'} = \mathrm{KL}(\mu_{a,x,s} \,\|\, \mu_{a,x,s'}) \ ,$$

for the KL-divergence between rewards under states s, s'. Our bounds and algorithms use a state s as reference point for the set of alternative states s' with different optimal arms,

$$\mathrm{Alt}_x(s) := \{s' : a^*_{x,s'} \neq a^*_{x,s}\} \ .$$

We can now derive the following result.

Lemma 1. *Given a problem instance with latent state s and observed context x, any δ-PAC algorithm ϕ must satisfy for any alternative state $s' \in \mathrm{Alt}_x(s)$,*

$$\sum_a \mathbb{E}_\phi[N_a \mid x, s]\mathbb{KL}^{R,a,x}_{s,s'} + \rho(x; s, s') \geq \mathbf{kl}(\delta||1-\delta), \tag{2}$$

where N_a is the number of plays of arm a drawn under ϕ and $\mathbf{kl}(\delta||1-\delta)$ is the KL-divergence between two Bernoulli random variables with parameters δ and $1-\delta$.

Proof summary. *The proof follows the argument of the original Lemma in Kaufmann et al. (2016). We start from the KL-divergence between the distribution of histories H, under s and s' and expand this using the chain-rule of the KL-divergence. We then apply the information-processing inequality to lower bound this by $\mathbf{kl}(\delta||1-\delta)$. The difference from Kaufmann et al. (2016) is that we get an additive term which depends on the context distribution under different latent models. For a full proof, see Appendix A.1.*

From lemma 1, we can derive a lower bound on the expected stopping time. Here, we assume that the optimal arm is unique for each state-context pair (s, x), that is, $\mathrm{Alt}_x(s) = \mathcal{S} \setminus \{s\}$. This assumption is *not* necessary to run our proposed algorithms.

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,

$$\begin{aligned} &\underset{\gamma_{x,a}\geq 0}{\text{minimize}} \ \sum_a \gamma_{x,a} \\ &\text{subject to} \ \sum_a \gamma_{x,a}\mathbb{KL}^{R,a,x}_{s,s'} + \frac{\rho(x;s,s')}{\mathbf{kl}(\delta||1-\delta)} \geq 1, \ \forall s' \in \mathrm{Alt}_x(s) \end{aligned} \tag{3}$$

Proof summary. *By lemma 1, we have a constraint on the sum of the expected number of times each arm is played by any δ-PAC algorithm ϕ. By dividing each side of equation 2 by $\mathbf{kl}(\delta||1-\delta)$ and minimizing the the stopping time under the resulting constraint, we obtain the linear program (LP) in equation 3. For a new bandit instance, x is observed before search begins. Thus, given a model $\mathcal{M}_\theta$, the only unknowns in equation 3 are $\gamma_{x,a}$. As we have a finite set of latent states s , we can construct a finite set of linear constraints and solve for the minimal stopping time. A full proof is given in appendix A.1.*

Remark 1. *As a sanity check, we verify that the contextual information makes the pure-exploration problem fundamentally easier. Indeed, when an observation x clearly separates the true latent state s from s', ρ increases, the constraint in equation 3 is satisfied by a larger set of parameters $\gamma_{x,a}$, and the lower bound attains a smaller value. However, as we require increasing certainty and $\delta \to 0$, the influence from contextual information X on $C^*_\delta(s,x)$ vanishes. This is expected since we don't collect more information through x as our requirement on certainty increases—it remains constant.*

As a consequence of proposition 1, we can obtain a bound for the population (marginal) search time. If we assume that $\frac{1}{C^*_\delta} = \mathbb{E}_{X,S}[\sum_a \gamma^*_{x,a}(s)]$ exists, with $\gamma^*_{x,a}$ the minimizers as in proposition 1, we have

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

The lower bound indicates that the optimal worst-case solution to equation 1 is limited by the hardest-to-separate states s, s'. We make use of this insight next to develop algorithms.

Algorithm 1 LLPT Explorer and Divergence Explorer

Input $\delta, T, \mathcal{S}, K, \mathcal{M}_\theta$
Output $\tau, \hat{i}_\tau$

Observe $h_1 = (x)$
if LLPT Explorer **then**
  Compute $w^*_{x,a}(s)$ for all a, s ▷ See equation 4, equation 3
end if

while $Z_t < 1 - \delta$ and $t < T$ **do**
  if LLPT Explorer **then**
    $s_t = \arg\max_{s \in \mathcal{S}} p(s|h_t)$
    $a_{t+1} = \arg\max_{a \in [K]} \; t \cdot w^*_{x,a}(s_t) - N_{a_t}(t)$
  else if Divergence Explorer **then**
    $s_t \sim p(s|h_t)$
    $f_t(a) = \sum_{s'} p_\theta(s'|h_t) \mathrm{KL}(\mu_{a,x,s_t} \,\|\, \mu_{a,x,s'})$
    $a_{t+1} = \arg\max_{a \in [K]} f_t(a)$
  end if
  Choose a_{t+1}, and **Observe** r_{t+1}
  Update $h_t = h_{t-1} \cup (a_{t+1}, r_{t+1})$
  Update $N_{a_{t+1}}(t) \leftarrow N_{a_{t+1}}(t) + 1$

  Update $\hat{s}_t = \arg\max_{s \in \mathcal{S}} p_\theta(s \mid h_t)$
  Update $\hat{a}_t = \arg\max_{a \in [K]} \mu_{a,x,\hat{s}_t}$
  Update $Z_t = \sum_s p_\theta(s|h_t) \mathbb{1}[\hat{a}_t = a^*_{x,s}]$
end while

Return $\hat{a}_t$

4 Algorithms

We present two best-arm identification strategies, each comprising a sampling rule for selecting arms A_t, a stopping rule for determining τ, and a recommendation rule for selecting $\hat{a}_\tau$. Both algorithms, defined in Algorithm 1, are given access to an *already estimated* latent variable model $\mathcal{M}_\theta$ including all reward means $\mu_{a,x,s} \;\forall\; s \in S, a \in A$ given a context x and differ only in their sampling rules; the stopping and recommendation rules are equivalent. Either algorithm starts by observing the random context X, and proceeds from there.

4.1 Sampling rule 1: Latent LP-based Track and Stop (LLPT) explorer

Our first sampling rule is based on the Track-and-Stop method (Garivier and Kaufmann, 2016), where arm allocations are determined by tracking proportions w^*, obtained by solving the lower bound optimization problem in equation 3. Since we have finite sets of states and actions, and x is observed at the start of the search, we can compute $\gamma^*_{x,a}(s)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ directly. Then, we define playing proportions $w^*_x(s)$, for each possible state $s \in \mathcal{S}$, as

$$w^*_{x,a}(s) = \gamma^*_{x,a}(s) / (\sum_a \gamma^*_{x,a}(s)) \; . \tag{4}$$

At each time step t, the algorithm picks a latent state $s_t = \arg\max_s p(s|h_t)$ from the (known) posterior given the current history h_t, and plays the arm which most closely tracks $w^*_{x,a}(s_t)$. Let $N_a(t)$ be the number of times arm a has been played up until and including t. Then, the *LLPT Explorer* sampling rule is defined by

$$A_{t+1} = \arg\max_{a \in [k]} \; t \cdot w^*_{x,a}(s_t) - N_a(t) \; .$$

The LLPT Explorer aims to play the minimum total number of trials using arms which distinguish latent states the most, as given by the KL term in the constraint of equation 3. It aims only to distinguish latent states with different optimal arms, as the goal is to identify the best action, not the state.

4.2 Sampling rule 2: Divergence explorer

The LLPT Explorer plays according to the optimal proportions for the worst-case alternative state given the current estimate. This is because the constraint in equation 3 will be hardest to satisfy (require largest $\gamma_{x,a}$) for states s' which are the most similar to s. A drawback of this idea is that it ignores the likelihood of said alternative state under the posterior. If there is strong evidence that s' is unlikely to be the true state, collecting more evidence to rule it out may be suboptimal. In the extreme case, a state s' with posterior probability $p(S = s' \mid h_t) \approx 0$ may still (unnecessarily) inform the sampling rule for the LLPT Explorer.

As an alternative, we define the *Divergence Explorer* sampling rule. This algorithm aims to play arms according to how much information is gained by playing an arm *in expectation* given the current posterior probability of states in $\mathrm{Alt}_x(s)$. At each time t, a latent state $s_t \sim P_t(s|h_t)$ is sampled as reference. Then, the sampling rule uses the expected divergence between s_t and alternative states s'_t,

$$f_t(s_t, a) = \sum_{s'_t \in \mathrm{Alt}_x(s_t)} P(s'_t|h_t)\mathrm{KL}(\mu_{a,x,s_t} \,\|\, \mu_{a,x,s'_t}) \ .$$

The arm $A_{t+1} = \arg\max_{a \in \mathcal{A}} f_t(s_t, a)$ is played next.

Because $\mathrm{KL}(\mu_{a,x,s_t} \,\|\, \mu_{a,x,s'_t})$ measures the information distance between the reward distribution of arm a under the two latent models s_t and s'_t, $f_t(s_t, a)$ does a one-to-many test assuming s_t is the true model and s'_t is another latent model with probability $P(s'_t|h_t)$.

4.3 Recommendation rule

Both algorithms recommend the best arm in the state most believed to be correct in a given instance, so the recommendation rule is $\hat{a}_\tau = \arg\max_{a \in \mathcal{A}} \mu_{a,x,\hat{s}_\tau}$ where $\hat{s}_\tau$ is the most probable state under the posterior, as defined in Algorithm 1.

4.4 Stopping rule

It is natural to stop search at t when we are confident enough that the recommended arm $\hat{a}_t$ is optimal under the posterior over latent states. Since we assume to have access to the full posterior over S, we can use the simple stopping rule

$$\tau := \min_t \{t : Z_t \geq 1 - \delta\} \quad \text{where} \quad Z_t = \sum_s P(s|h_t)\mathbb{1}[\hat{a}_t = a^*_{x,s}] \tag{5}$$

and the threshold $1 - \delta$ is the desired confidence level. Whenever this rule is satisfied, so is Chernoff's stopping rule based on a threshold $\log(\frac{1-\delta}{\delta})$ on the log-likelihood ratio between states, as used by Garivier and Kaufmann (2016). See the proof of proposition 2 in appendix A.2 for a derivation.

In many applications, it us sufficient to identify a action which is ϵ-optimal with respect to the best possible action in the true latent state. We can accommodate this in our algorithm by redefining the set of alternative states s' to include only those for which the optimal arm in s is more than ϵ worse than the optimal arm in s',

$$\mathrm{Alt}_x(s) \coloneqq \{s' : \mu_{a^*_{x,s},x,s'} < \mu^*_{x,s'} - \epsilon\} \ .$$

This change involves only a minor modification to the stopping criterion in equation 5 and could also be used in the Divergence explorer sampling rule.

5 Upper bound on the expected stopping time of LLPT explorer

Next, we show that the lower bound derived in Section 3 is matched by an upper bound on the stopping time for the LLPT Explorer algorithm in the high-confidence limit, $\delta \to 0$. Similar to the lower bound, we make

the simplifying assumption that each latent state has a unique optimal arm, shared with no other states, $\text{Alt}_x(s) = \mathcal{S} \setminus \{s\}$. As a consequence, finding the optimal arm equates to finding the true underlying state. We have the following result.

Proposition 2. *Let τ be the stopping time of LLPT Explorer ϕ, as defined in Algorithm 1. With s the true state and $C^*(s, x)$ the optimum in equation 3 with the ρ-term removed, there is a constant $\alpha > 0$ such that*

$$\lim_{\delta \to 0} \frac{\mathbb{E}_\phi[\tau \mid s, x]}{\log(1/\delta)} \leq \frac{\alpha}{C^*(s, x)} \ . \tag{6}$$

Proof summary. *The proof combines and expands arguments from Garivier and Kaufmann (2016) and Chernoff (1959) to show that after sufficently many samples, a) the true latent state is identified, b) the tracked proportions are near optimal for the identified state, c) the probability that the stopping criterion is not satisfied decays exponentially quickly. As a result, the expected stopping time can be bounded using concentration arguments. For a proof, see appendix A.2.* ■

As stated, proposition 2 applies to the LLPT Explorer, as defined in Algorithm 1, in which the MAP state $\hat{s}_t$ is used for tracking. We have also implemented a slight variation of LLPT with a sampled state $\hat{s}_t \sim p(s|h_t)$ and found that the latter worked slightly better empirically. We report only results for the version in Algorithm 1.

Similarly to the lower bound, we obtain an upper bound on the population search time by taking the expectation of equation 6 with respect to S and X.

Remark 2. *Comparing the result in equation 6 to bounds for pure-exploration without latent variable models; e.g., Russo (2016); Garivier and Kaufmann (2016), superficially, they appear very similar. However, the critical quantity in the classical setting is the smallest separation of reward means for alternative, free vectors of arm parameters. Here, the equivalent quantity is the set of parameters of the discrete latent states, which is generally much smaller than the set of free parameters, leading to a tighter bound.*

More precisely, 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 our case, $\text{Alt}_x(s)$ 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.

6 Experimental study

We evaluate our proposed algorithms in a series of experiments, comparing them to baseline algorithms for fixed-confidence pure exploration.

6.1 Baseline algorithms

Previous work incorporating latent states in pure exploration was not available at the time of writing, so to get comparable baselines, we adapted the Top-Two Thompson Sampling (TTTS) rule (Russo, 2016) to compare to our algorithms.

Top-Two Thompson Sampling (TTTS) TTTS operates with the goal of estimating parameters Π_t (e.g., mean vectors of arms with Gaussian distribution) that yield the best arm for a given confidence level $1 - \delta$. It proceeds as follows; at each time step t either; (i) with probability p, sample a parameter vector $\theta_t \sim \Pi_t$ and play the arm ${a_t}^{(1)} = \arg\max_{a \in \mathcal{A}} \theta_t$ or (ii) with probability $1 - p$ resample ${\theta_t}' \sim \Pi_t$ until it gets and subsequently plays arm ${a_t}^{(2)} \neq {a_t}^{(1)}$. We implemented the T3C (Shang et al., 2020) variant of TTTS which finds ${a_t}^{(2)} \neq {a_t}^{(1)}$ faster. TTTS does not make use of a latent variable model.

TTTS-Latent Explorer This is an adaptation of TTTS to our setting where, instead of estimating arm parameters, the goal is purely to identify the latent state. It does not account for the case where there is a shared optimal arm over different states which is accounted for in the LLPT and Divergence Explorer.

At each time step t, the sampling rule samples a latent state $s_t^{(1)} \sim P_t(s|h_t)$ and either (i) with a Bernoulli parameter p evaluates the latent state, $s_t = s_t^{(1)}$ or (ii) with a Bernoulli parameter $1-p$ resamples $P_t(s|h_t)$ until it gets a latent state $s_t = s_t^{(2)} \neq s_t^{(1)}$. It then plays the arm $A_t = \arg\max_{a \in \mathcal{A}} \mu_{a,x,s_t}$.

Greedy Explorer This is a naïve sampling rule which plays the reward-optimal arm in a state sampled from the current posterior, akin to TTTS-Latent but without the re-sampling step. At each time t, it picks $s_t = \arg\max_s p(s|h_t)$ and then plays the locally reward-maximizing arm $A_t = \arg\max_{a \in \mathcal{A}} \mu_{a,x,s_t}$. It is naïve in the sense that it only considers the rewards from a state, but this is not always informative for distinguishing alternative states. It also corresponds to standard Thompson Sampling (Thompson, 1933) which has been shown to perform poorly for pure exploration tasks, hence the motivation for TTTS.

6.2 Experimental environment

As treatment personalization task, we use the Alzheimer's Disease Causal estimation Benchmark (ADCB) environment (Kinyanjui and Johansson, 2022). In this environment, simulated subjects go through cognitive decline, eventually progressing into Alzheimer's disease. Outcomes Y_t represent their cognitive abilities and treatments A_t are symptomatic, affecting only immediate outcomes. Both treatment responses and an initial 33-dimensional observed context $X \in \mathbb{R}^d$, are affected by a latent state S, representing the disease subtype.

In the ADCB environment, the number of actions is $K = 8$ and the number of latent states, $S = 6$. The outcome Y_t at time t is generated as $Y_t(A, X, S) := \Phi(X, S) + \Delta(A_t, S) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$ and Φ is an non-linear function fit to real data to model the cognitive function of subjects when not treated. For the environment we are using, Φ is a Random Forest Regressor fit to observed outcomes of untreated patients. Δ is a function that is defined to moderate the heterogeneity of simulated treatment effects over the latent dimensions. Here, $\Delta := \upsilon \mathbb{1}_S + \mathbb{1}_S(\eta \upsilon \beta^T)$ where $\upsilon \in \mathbb{R}^K$ is the average treatment effect of the treatments, $\eta > 0$ is a heteregoneity scaling parameter, and $\beta \in \mathbb{R}^{K \times S}$ is a factor matrix whose rows sum to 0.

We define two alternative reward settings (see below), both with Gaussian rewards, based on the ADCB outcomes of treatments, Y. We give algorithms which make use of latent variables perfect knowledge of the true latent variable model, as defined by the simulator. Hence, for each context $x \in \mathbb{R}^d$, latent state $s \in [S]$ and action $a \in [K]$, the corresponding posterior $p(s \mid h_t)$ and reward means, $\mu_{a,x,s}$ are known.

Reward setting 1: Non-contextual rewards Here, for each latent state we define the reward $R := -(Y(A, X, S) - Y(0, X, S))$. From the definition of the outcome Y above, this removes the effect of context from the reward, by cancelling $\Phi(X, s)$, and takes us closer to a typical best arm identification setting with additional latent state structure, where the structure is given by Δ. In appendix B.1, Figure 5(a) shows the structure of the mean rewards $\mu_{a,x,s}$ under the different latent states $s \in \mathcal{S}$, $a \in K$ for this setting.

Reward setting 2: Contextual rewards Here, we define the reward $R := -Y(A, X, S)$, thus preserving the effect of context in the reward. As seen from appendix B.1, Figure 5(b), which is an example of the mean rewards structure $\mu_{a,x,s}$ $s \in \mathcal{S}$, $a \in K$ for some given context x, the reward structure stays the same as in the previous setting, but the scale is shifted depending on the context. The similarity is a property of the environment. The results presented in the results section below are for this setting, and those of setting 1 above are appended in the supplementary materials.

Repeated experiments Each experiment proceeds as follows; A new patient is sampled from the environment (sampled patients have potentially different latent states and contexts). The algorithms do not observe the latent state and they proceed as described in Section 4 and Section 6.1. For a run, all algorithms are provided with the same context. All results are presented for 100 different patients and averages are computed for the different quantities compared. Errorbars represent the standard deviation across patients.

Evaluation metrics We compare empirical estimates of the expected stopping time $\mathbb{E}[\tau]$, convergence of the posterior probability $p(\hat{s}_t \mid h_t)$ with t, and the average correctness level, $\mathbb{E}[\mathbb{1}[\hat{a}_\tau = a^*]]$, of the different algorithms for i) different levels of confidence $\delta \in (0, 1/2)$ under a fixed noise level $\sigma > 0$ and ii) different levels

of noise σ for a fixed δ. Results for correctness are presented in Figure 6 in the Appendix, and correspond closely with the parameter δ.

6.3 Results

In Figure 2, we see an example of the drastic effect that incorporating latent structure can have on the stopping time of pure-exploration algorithms. All latent-variable methods outperform the non-latent baseline TTTS by a substantial margin.

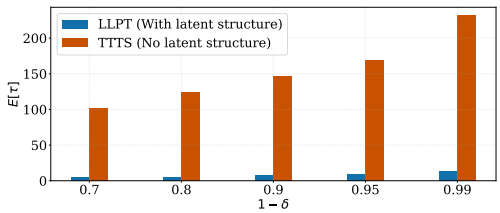


Figure 2: Using latent state structural information significantly reduces the expected number of trials $\mathbb{E}[\tau]$ required to identify an optimal treatment with confidence at least $1 - \delta$ in a simulator of Alzheimer's disease progression.

Moreover, in Figure 3a, we see that, even for the worst-case instances, the LLPT algorithm is faster than the average for standard TTTS observed in Figure 2. This supports our hypothesis that exploiting latent structure between instances (patients), which could be estimated from historical data, contexts, is useful to design sample-efficient pure-exploration algorithms.

In the graph of latent state posterior convergence, Figure 3b, we see that LLPT Explorer and Divergence Explorer converge quicker in their belief of the inferred latent state. We also observe less variance across bandit instances (shaded area) compared to the Greedy and TTTS-Latent baselines. The implication for this is that these algorithms stop exploration earlier thus attaining our goals outlined in Section 2.

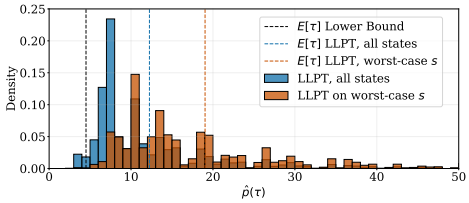
In Figure 3c, we study the average stopping time, $\hat{\mathbb{E}}[\tau]$ for all algorithms with access to the same latent variable model, under changing certainty level $1 - \delta$. LLPT Explorer and Divergence Explorer are consistently more efficient than baselines, demonstrating benefit of the insights derived from the lower bound in proposition 1. The difference is especially pronounced in the high-certainty regime, $\delta \approx 0$, which is the regime that would be ideal for safety-critical healthcare applications. Interestingly, we find that the Divergence Explorer performs consistently better than the LLPT Explorer and its average stopping time approaches the lower bound as $\delta \to 0$. We believe this is due to selecting actions based on comparison with alternative states on average under the current posterior, rather than the worst-case alternative state - some latent states are ruled out by the posterior and no longer affect the action selection of the divergence explorer.

Studying our algorithms with respect to noise in the rewards, Figure 3d, shows that our proposed methods are also more robust to noise compared to the baseline algorithms. At $\sigma = 10$, which is comparable to the marginal standard deviation of rewards due to X and S, we see that our algorithms perform better. We also observe that they are also more robust to over- and under-estimation of the noise level in the rewards as shown by $\mathbb{E}[\tau]$ at other noise levels.

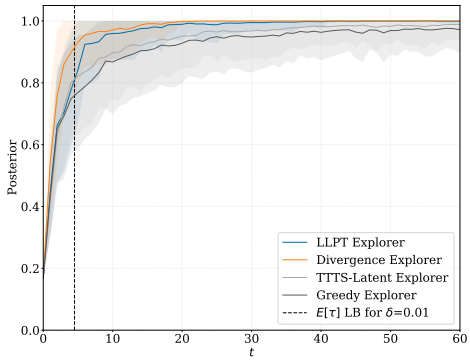
7 Related work

The problem of finding optimal decisions under uncertainty has a long history (Thompson, 1933; Chernoff, 1959; Gittens and Dempster, 1979; Jennison et al., 1982; Lai and Robbins, 1985; Glynn and Juneja, 2004) and has recently been studied as a pure exploration problem in the multi-armed bandit framework under

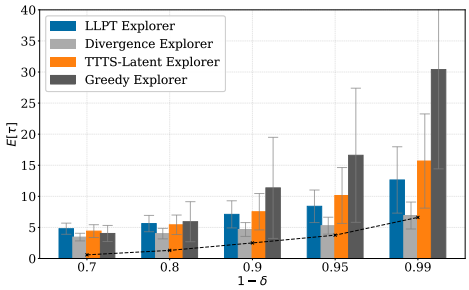
[2]The small discrepancy seen in the case where $\sigma = 1$ is due to the exclusion of the ρ term in the computed lower bound.



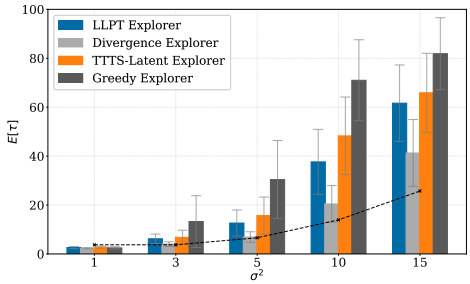
(a) Density of stopping times under LLPT(ours) showing worst-case latent state ($\delta = 0.01$, Number of patients, $N = 10,000$). The variance of the stopping time under all the latent states is reasonably low. The higher stopping times can be attributed to the worst-case latent states, though they are still reasonably low.



(b) Comparison of posterior convergence of the different algorithms [$\delta = 0.01$, Number of patients, $N = 100$]. The posteriors for our algorithms, LLPT Explorer and Divergence Explorer, concentrate more quickly.



(c) Comparison of stopping time vs confidence ($1 - \delta$) for the algorithms. Our algorithms, LLPT Explorer and Divergence Explorer, have stopping times that are consistently lower. The dashed line shows the lower bound from Proposition 1.



(d) Comparison of stopping time vs noise for the algorithms Our algorithms, LLPT Explorer and Divergence Explorer, are consistently more robust to noisy rewards compared to the baselines. The dashed line shows the lower bound from Proposition 1.[2]

Figure 3: Selected results from our experimental study.

various assumptions(Even-Dar et al., 2006; Bubeck et al., 2009; Jamieson et al., 2013; Kaufmann et al., 2016; Garivier and Kaufmann, 2016; Jedra and Proutiere, 2020; Wang et al., 2021; Agrawal et al., 2021; Tirinzoni and Degenne, 2022).

The work of Garivier and Kaufmann (2016) is the first to introduce an optimal algorithm, Track and Stop, in the fixed confidence setting for classical multi-armed bandits and our LLPT Explorer takes inspiration from their algorithm, adapting it to the latent bandit setting. Russo (2016) introduces a class of top-two sampling strategies for the pure-exploration problem, which we here use as baselines. These top-two algorithms were originally analyzed using a different performance measure but have recently been theoretically analyzed in the fixed-confidence setting by Jourdan et al. (2022). Our work is also related to (Maillard and Mannor, 2014; Zhou and Brunskill, 2016; Hong et al., 2020a;b), who study regret minimization in latent bandits, in contrast to our work which studies the pure-exploration problem in latent bandits.

Kato and Ariu (2021) studied pure exploration in contextual bandits, where a new context is observed at each time point, and found that contextual information improves the speed at which the average treatment

effects (Imbens and Rubin, 2015) of actions across contexts can be estimated. Our problem is related to this setting but differs in that we see only a single context x per bandit instance, and are interested in the effects of actions for this specific x, not on average. Håkansson et al. (2020) studied fast search for near-optimal treatments, based on a model learned from historical trajectories, but did not consider online learning. In their setting, an optimal search strategy can be found by solving a dynamic programming problem in an estimated discrete state space. This is not feasible here due to the high dimensionality of our history, H.

8 Discussion & conclusion

In this work we have studied the problem of finding the optimal arm in latent bandits using as few trials as possible. We have empirically and theoretically shown that our proposed algorithms are able to leverage the latent structure in a near-optimal way to substantially reduce the expected stopping time compared to available baselines. Our empirical evaluation in a simulator of Alzheimer's disease derived from real-world data, demonstrated that our algorithms are able to find the optimal treatment in just a few trials.

Our analysis is limited to the case in which the latent variable model is given and exact. When forced to estimate the model from historical data, sensitivity to misspecification or misestimation becomes a concern. Hong et al. (2020a) analyzed latent bandits in regret minimization when the reward model is misspecified but the resulting bound suffers linear regret scaled by the error, and Hong et al. (2022) provided an improved sub-linear regret bound for this with additional assumptions on the reward structure. In the pure-exploration setting, recovering quickly from misspecification is even more critical since the time scale is shorter. We conjecture that an informative guarantee in the misspecified case will similarly require additional assumptions on the reward structure or additional sources of data. We believe the setting where a learner needs to recover the true model up to some pre-specified precision is an interesting direction for future work. Another useful generalization would be to go beyond the analysis of expected rewards. In high-stakes applications, it is desirable to manage also the risk of worst-case low-probability events, see e.g., Tamkin et al. (2019). This would further increase the suitability of our approach for the medical domain.

References

Shubhada Agrawal, Wouter M Koolen, and Sandeep Juneja. Optimal best-arm identification methods for tail-risk measures. In *Advances in Neural Information Processing Systems*, volume 34, pages 25578–25590, 2021.

Larry Borish and Jeffrey A Culp. Asthma: a syndrome composed of heterogeneous diseases. *Annals of Allergy, Asthma & Immunology*, 101(1):1–9, 2008.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer, 2009.

Bibhas Chakraborty and EE Moodie. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959. ISSN 00034851.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Gayatri Devi and Philip Scheltens. Heterogeneity of alzheimer's disease: consequence for drug trials? *Alzheimer's Research & Therapy*, 10(1):1–3, 2018.

M Dragomirescu and C Bergthaller. On the continuity of the optimum of a linear program. *Studii si Cercetari Mathematice*, 18:1197–1200, 1966.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39): 1079–1105, 2006.

Paul V Fish, David Steadman, Elliott D Bayle, and Paul Whiting. New approaches for the treatment of alzheimer's disease. *Bioorganic & medicinal chemistry letters*, 29(2):125–133, 2019.

Liana Fraenkel, Joan M Bathon, Bryant R England, E William St. Clair, Thurayya Arayssi, Kristine Carandang, Kevin D Deane, Mark Genovese, Kent Kwas Huston, Gail Kerr, et al. 2021 american college of rheumatology guideline for the treatment of rheumatoid arthritis. *Arthritis & Rheumatology*, 73(7): 1108–1123, 2021.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

J. Gittens and Michael Dempster. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B: Methodological*, 41:148–177, 02 1979.

Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *In Proceedings of the 2004 winter simulation conference*, volume 1, pages 577–585, 01 2004.

Samuel Håkansson, Viktor Lindblom, Omer Gottesman, and Fredrik D Johansson. Learning to search efficiently for causally near-optimal treatments. *Advances in Neural Information Processing Systems*, 33: 1333–1344, 2020.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13423–13433. Curran Associates, Inc., 2020a.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Mohammad Ghavamzadeh, and Craig Boutilier. Non-stationary latent bandits. *CoRR*, abs/2012.00386, 2020b.

Joey Hong, Branislav Kveton, Manzil Zaheer, Mohammad Ghavamzadeh, and Craig Boutilier. Thompson sampling with a mixture prior. In *International Conference on Artificial Intelligence and Statistics*, pages 7565–7586. PMLR, 2022.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. *Journal of Machine Learning Research*, 35, 12 2013.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10007–10017. Curran Associates, Inc., 2020.

Christopher Jennison, Iain M Johnstone, and Bruce W Turnbull. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical decision theory and related topics III*, pages 55–86. Elsevier, 1982.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. *arXiv preprint arXiv:2206.05979*, 2022.

Masahiro Kato and Kaito Ariu. The role of contextual information in best arm identification. *arXiv preprint arXiv:2106.14077*, 2021.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42, jan 2016. ISSN 1532-4435.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Newton Mwai Kinyanjui and Fredrik D Johansson. Adcb: An alzheimer's disease simulator for benchmarking observational estimators of causal effects. In *Conference on Health, Inference, and Learning*, pages 103–118. PMLR, 2022.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34:24523–24534, 2021.

T.L Lai and H Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. *31st International Conference on Machine Learning, ICML 2014*, 05 2014.

Daniel Russo. Simple bayesian algorithms for best arm identification. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1417–1418, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Xuedong Shang, Rianne Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR, 2020.

John M Stern. Overview of evaluation and treatment guidelines for epilepsy. *Current treatment options in neurology*, 11(4):273–284, 2009.

Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

Joy A. Thomas Thomas M. Cover. *Entropy, Relative Entropy, and Mutual Information*, pages 13–55. John Wiley & Sons, Ltd, 2005.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444.

Andrea Tirinzoni and Rémy Degenne. On elimination strategies for bandit fixed-confidence identification. *arXiv e-prints*, pages arXiv–2205, 2022.

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

Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743*, 2016.

ANGELA Zink, JOACHIM Listing, SABINE Ziemer, HENNING Zeidler, German Collaborative Arthritis Centres, et al. Practice variation in the treatment of rheumatoid arthritis among german rheumatologists. *The Journal of rheumatology*, 28(10):2201–2208, 2001.

Appendix

A Proofs

Our objective can be written as follows

$$\begin{aligned} &\underset{\phi}{\text{minimize}} && \mathbb{E}_{H_\tau, S, \phi}[\tau] \\ &\text{subject to} && P(\mu_{\hat{a}_\tau, x, s} < \ \mu^*_{x,s} \mid X = x, S = s) \leq \delta, \ \ \forall x, s \end{aligned} \tag{7}$$

A.1 Lower bound

Recall the definition of $\text{Alt}_x(s)$, given a latent state s we define the set of alternative latent states as

$$\text{Alt}_x(s) := \{s' \in \mathcal{S} : \arg\max_a \mathbb{E}[r|s, x, a] \neq \arg\max_a \mathbb{E}[r|s', x, a]\}. \tag{8}$$

Proof of lemma 1

Recall the statement of lemma 1, *Given a latent state s and context x, any δ-PAC algorithm ϕ will satisfy*

$$\sum_a \mathbb{E}_\phi[N_a|x, s]\mathbb{KL}^{R,a,x}_{s,s'} + \rho(x; s, s') \geq \mathbf{kl}(\delta||1-\delta). \tag{9}$$

Proof. Let H_t denote the history up to time t. The expected log-ratio between s and $s' \in \text{Alt}_x(s)$ under the latent state s and algorithm ϕ can be written as

$$\begin{aligned} \mathbb{E}_\phi[L_t(s, s')|x, s] &= \mathbb{E}_\phi\left[\log \frac{p(H_t|s)}{p(H_t|s')}|x, s\right] && (10) \\ &= \mathbb{E}_\phi\left[\rho(x; s, s') + \sum_{i=1}^{t} \log \frac{p(r_i|s, a_t, x)}{p(r_i|s', a_t, x)}|x, s\right] && (11) \\ &= \rho(x; s, s') + \sum_{a=1}^{K} \mathbb{E}_\phi[N_a|x, s]\mathbb{KL}^{R,a,x}_{s,s'} && (12) \end{aligned}$$

where the last step follows from the KL-divergence decomposition, see Lemma 15.1 in (Lattimore and Szepesvári, 2020). Further, by definition we have

$$\text{KL}(p_\phi(H_t|x, s) \,\|\, p_\phi(H_t|x, s')) = \mathbb{E}_\phi[L_t(s, s')|x, s] \tag{13}$$

and using the information-processing inequality (Thomas M. Cover, 2005), as in (Kaufmann et al., 2016) yields

$$\mathbb{E}_\phi[L(s, s')|x, s] \geq \mathbf{kl}(\delta||1-\delta) \tag{14}$$

where $\mathbf{kl}(\delta||1-\delta)$ is the KL-divergence between two Bernoulli variables with mean δ and $1-\delta$. ■

Proof of proposition 1

Proof. This proof follows the same line as the proof for the general lower bound in (Kaufmann et al., 2016). The main difference is that we, due to lemma 1, get a dependence on the context distribution, $p(X|s)$, in the lower bound.

From lemma 1 we have

$$\rho(x; s, s') + \sum_{a=1}^{K} \mathbb{E}_\phi[N_a|x, s]\mathbb{KL}^{R,a,x}_{s,s'} \geq \mathbf{kl}(\delta||1-\delta), \forall x \text{ and } \forall s' \in \text{Alt}_x(s). \tag{15}$$

Equation 15 gives a necessary condition which any δ-PAC algorithm needs to obey and we can simply minimize $\mathbb{E}_\phi[\tau|x,s]$ w.r.t. this constraint. Note that this yields a LP with finite constraints since the set of all latent states is finite. Hence, we get the following optimization problem

$$\begin{aligned} &\underset{\phi}{\text{minimize}} && \mathbb{E}_\phi[\tau|x,s] \\ &\text{subject to} && \sum_{a=1}^{K} \mathbb{E}_\phi[N_a|x,s]\mathbb{KL}^{R,a,x}_{s,s'} + \rho(x:s,s') \geq \mathbf{kl}(\delta||1-\delta); \quad \forall s' \in \text{Alt}_x(s) \end{aligned}$$

We introduce

$$\gamma_{x,a} := \frac{\mathbb{E}[N_a|x,s]}{\mathbf{kl}(\delta||1-\delta)} \tag{16}$$

and solving the above optimization problem is equivalent to solving

$$\begin{aligned} &\underset{\gamma_{x,a}\geq 0}{\text{minimize}} && \sum_a \gamma_{x,a} \\ &\text{subject to} && \sum_a \gamma_{x,a}\mathbb{KL}^{R,a,x}_{s,s'} + \frac{\rho(x;s,s')}{\mathbf{kl}(\delta||1-\delta)} \geq 1, \quad \forall s' \in \text{Alt}_x(s). \end{aligned} \tag{17}$$

Let $\gamma^*_{x,a}$ be a optimal solution, then

$$\mathbb{E}[\tau|x,s] = \sum_a \mathbb{E}[N_a|x,s] \geq \mathbf{kl}(\delta||1-\delta) \sum_a \gamma^*_{x,a} \tag{18}$$

and by defining $1/C_\delta(s,x) = \sum_a \gamma_{x,a}$ we get

$$\mathbb{E}[\tau|x,s] \geq \mathbf{kl}(\delta||1-\delta)\frac{1}{C_\delta(s,x)}. \tag{19}$$

■

A.2 Upper bound on sample complexity for tracking rule

Let τ represent the (random) stopping time with certainty parameter δ. Further, let $L_t(s,s')$ represent the log-likelihood ratio of t samples under model s and s',

$$L_t(s,s') = \rho(x_i;s,s') + \sum_{i=1}^{t} z_i(s,s') \;\text{ where }\; \sum_{i=1}^{t} z_i(s,s') := \log\frac{p(r_i \mid S=s, A=a_i)}{p(r_i \mid S=s', A=a_i)} \tag{20}$$

and

$$\rho(x_i;s,s') = \log\frac{p(x_i \mid S=s)}{p(x_i \mid S=s')}\ .$$

Next, let the optimal worst-case playing proportions $w^*_{x,a}(s) = \gamma^*_{x,a}/\sum_b \gamma^*_{x,b}$ in an observed context x under an assumed true state s be given by the optimizers $\gamma^*_{x,a}$ of equation 17.

When the context X is constant, the second term in the constraint vanishes and the $\gamma_{x,a}$ parameters is independent of x.

Proposition. *The LLPT algorithm ϕ (Algorithm 1) which a) selects actions by tracking proportions $w^*_{a,x}(\hat{s}_t) \propto \gamma^*_{a,x}(\hat{s}_t)$, where $\gamma^*_{a,x}(\hat{s}_t)$ are the solution to equation 17 with $\delta = 0$ and $\hat{s}_t$ is the MAP state at time t, and b) stops according to the stopping rule in Section 4.4, satisfies, with s the true state, and a constant $\alpha > 0$,*

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

Proof. We make an adaptation of the proof of Lemma 2 in (Chernoff, 1959) to tracking algorithms with an initial observed context. Let a_i be actions drawn according to a tracking rule which selects actions according to a concentrating parameter (in our case $\hat{s}_t$ concentrates to s and we track $w^*_{a,x}(\hat{s}_t)$), and let $N_a(t) = \sum_{i=1}^t \mathbb{1}[a_i = a]$. Then, by Lemma 17 in (Garivier and Kaufmann, 2016), for any $\zeta := \zeta_x(s)$, there exists a T_ζ such that for $T \geq T_\zeta$, we have

$$\forall t \geq \sqrt{T} : \max_a \left| \frac{N_a(t)}{t} - w^*_{x,a}(s) \right| \leq 3(K-1)\zeta \ .$$

Now, let $T_0 = \inf_t\{t : \forall t' \geq t, \hat{s}_{t'} = s\}$ be the smallest number of samples such that for more samples, the estimated latent state will be correct. This bound exists, and is reached exponentially fast, by Lemma 1 in (Chernoff, 1959):

$$p(T_0 > t) \leq Ke^{-bt} \ .$$

Next, let $T_{s'}(\delta) = \inf_t\{t : \forall t' \geq t, L_{t'}(s, s') > \log(\frac{1-\delta}{\delta})\}$ be the shortest time after the log-likelihood ratio exceeds $\log(\frac{1-\delta}{\delta})$ w.r.t. comparison between s and s'. Whenever the stopping criterion in Section 4.4 is satisfied with parameter δ, so is this. We can see this by noting that if $p(S = s \mid h_t) > 1 - \delta$ for some s, then $p(S = s' \mid h_t) < \delta$ for $s' \neq s$. Hence,

$$\log \frac{p(S = s \mid h_t)}{p(S = s' \mid h_t)} = L_t(s, s') > \log\left(\frac{1-\delta}{\delta}\right) \ .$$

It follows that,

$$\tau \leq \max(\max_{s' \neq s} T_{s'}(\delta), T_0, T_\zeta) \ .$$

We have from lemma 1 in (Chernoff, 1959) that there exist constants K and b such that

$$p(T_0 > t) \leq Ke^{-bt} \ .$$

Hence, to show that the stopping time is bounded by t, it is sufficient to show that for each alternative state $s' \neq s$, and sufficiently large t, there are constants $K = K(\epsilon, s'), b = b(\epsilon, s')$, such that

$$p(T_{s'}(\delta) > t) \leq Ke^{-bt} \ .$$

If the result holds for $t > \alpha \log(\frac{1-\delta}{\delta})/C^*_\delta(s, x)$, we have our result by a simple argument.

For $\zeta > 0$, define $W^\zeta = \{w{:=}\ w_{x,a}(s) \in [0,1]^K : \|w\|_1 = 1, \|w - w^*_{x,a}(s)\|_\infty \leq 3(K-1)\zeta\}$ to be the set of playing proportions ζ-close to $w^*_{x,a}(s)$. Now, define the ζ-worst-case playing proportions $w^\zeta(s)$ as the optimizers of $C^\zeta(s, x) = \min_{w \in W^\zeta} \min_{s'} \sum_a w_{x,a} \mathbb{KL}(\mu_{a,x,s}, \mu_{a,x,s'})$.

Consider $L_t(s, s')$ as defined in equation 20. Add and subtract both $\mathbb{KL}^{R,a_i,x}_{s,s'} := \mathbb{KL}(\mu_{a_i,x,s}, \mu_{a_i,x,s'})$ and $\mathbb{KL}^{R,w^\zeta,x}_{s,s'} := \mathbb{E}_{a \sim w^\zeta(s)}[\mathbb{KL}(\mu_{a,x,s}, \mu_{a,x,s'})]$ from term i in the sum,

$$\begin{aligned} L_t(s, s') &= \sum_{i=1}^t \left[z_i(s, s') - \mathbb{KL}^{R,a_i,x}_{s,s'} + \mathbb{KL}^{R,a_i,x}_{s,s'} - \mathbb{KL}^{R,w^\zeta,x}_{s,s'} + \mathbb{KL}^{R,w^\zeta,x}_{s,s'} \right] + \rho(x; s, s') \\ &= \underbrace{\sum_{i=1}^t \left[z_i(s, s') - \mathbb{KL}^{R,a_i,x}_{s,s'} \right]}_{(a)} + \underbrace{\sum_{i=1}^t \left[\mathbb{KL}^{R,a_i,x}_{s,s'} - \mathbb{KL}^{R,w^\zeta,x}_{s,s'} \right]}_{(b)} + \underbrace{t\mathbb{KL}^{R,w^\zeta,x}_{s,s'}}_{(c)} + \underbrace{\rho(x; s, s')}_{(d)} \ . \end{aligned}$$

Starting with term (a), by definition, for any time point i, by definition of the KL-divergence,

$$\mathbb{E}[z_i(s, s')] = \mathbb{E}_R \left[\log \frac{p(R \mid S = s, X = x, A = a_i)}{p(R \mid S = s', X = x, A = a_i)} \mid S = s \right] = \mathbb{KL}^{R,a_i,x}_{s,s'} \ .$$

Hence, for any $\epsilon_1 > 0$, $(\sum_{i=1}^{t}[z_i(s,s') - \mathbb{KL}_{s,s'}^{R,a_i,x}] + \epsilon_1)$ has positive mean and finite moment generating function for moments $k \in [-1, 0]$ for any a_i and $s' \neq s$. As a result, there exists $k^* < 0$ and $b_1 > 0$, depending on ϵ_1, such that for any trial i,

$$\mathbb{E}[e^{k^*[z_i(s,s') - \mathbb{KL}_{s,s'}^{R,a_i,x} + \epsilon_1]}] \leq e^{-b_1} \ .$$

Following the proof of Lemma 1 in (Chernoff, 1959), we have,

$$\mathbb{E}\left[e^{k^*[\sum_{i=1}^{t}[z_i(s,s') - \mathbb{KL}_{s,s'}^{R,a_i,x} + \epsilon_1]}\right] \leq e^{-b_1 t}$$

and, as a result,

$$p\left(\sum_{i=1}^{t}[z_i(s,s') - \mathbb{KL}_{s,s'}^{R,a_i,x}] < -\epsilon_1 t\right) \leq e^{-b_1 t} \ .$$

For term (b), it follows from the definition of w^ζ, T_ζ and C^ζ that, for any $t \geq \max(T_\zeta, T_0)$, $\hat{s}_t = s$ and $\|w(t) - w^*(s)\|_\infty \leq 3(K-1)\zeta$. Hence,

$$\sum_{i=1}^{t}[\mathbb{KL}_{s,s'}^{R,a_i,x} - \mathbb{KL}_{s,s'}^{R,w^\zeta,x}] \geq 0 \ .$$

In other words, after we have collected more than T_ζ samples, we will have more information than the ζ-worst-case rule for s. For term (c), by definition of C^ζ, for any s', $\mathbb{KL}_{s,s'}^{R,w^\zeta,x} \geq C^\zeta(s,x)$.

Combining the previous results, noting that term (d) is a constant, for any s' and any $\epsilon_3 > 0$ and appropriately chosen K_4, b_4, we get that for $t \geq \max(T_0, T_\zeta)$,

$$p\left(L_t(s,s') < t[C^\zeta(s,x) - \epsilon_3]\right) \leq K_4 e^{-b_4 t} \ .$$

For $t > \log(\frac{1-\delta}{\delta})/(C^\zeta(s,x) - \epsilon_3)$, we thus have

$$p(T_{s'} > t) \leq K_4 e^{-b_4 t} \ .$$

For any positive random variable T, we have the identity,

$$\mathbb{E}[T] = \int_0^\infty p(T > t)dt \ .$$

Hence,

$$\mathbb{E}[T_{s'}] \leq t_0 + \int_0^\infty p(T'_s > t)dt \leq t_0 + K_4/b_4$$

and so we can let $t_0 \geq T_0 + T_\zeta + \log(\frac{1-\delta}{\delta})/(C^\zeta(s) - \epsilon_3) + K_4/b_4$.

Next, we study the high-certainty limit $\delta \to 0$. We note first that as $\delta \to 0$, $\log(\frac{1-\delta}{\delta}) \to \log(1/\delta)$. When $\delta \to 0$, the influence of the term $\rho(x; s, s')$ in equation 3 vanishes and the $C_\delta^*(s,x)$ converges to

$$\begin{aligned} C^*(s,x) = \min_{\gamma_{x,a} \geq 0} & \sum_a \gamma_{x,a} \\ \text{s.t.} & \sum_a \gamma_{x,a} \mathbb{KL}_{s,s'}^{R,a,x} \geq 1, \ \forall s' \in \text{Alt}_x(s) \end{aligned} \tag{21}$$

by the continuity of linear programs (Dragomirescu and Bergthaller, 1966). Thus, if we let $\zeta \to 0$, we have $C^\zeta(s,x) \to C^*(s,x)$. We get,

$$\lim_{\delta \to 0} \frac{\mathbb{E}[\tau \mid x]}{\log 1/\delta} \leq \frac{1}{(C^*(s,x) - \epsilon_3)} \ .$$

Refactoring, we get the desired result.

■

B Additional experiments and results

B.1 Reward Structure

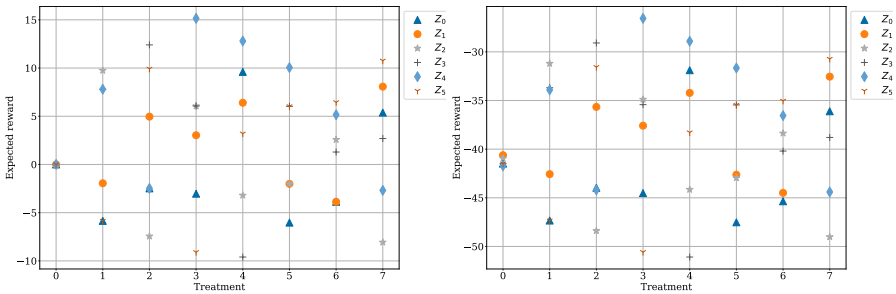


Figure 4: Structure of the means $\mu_{s,a}$ under different latent states. (a) Non-contextual rewards and (b) Contextual rewards

B.2 Outcome Distribution

Shown in Figure 5 below.

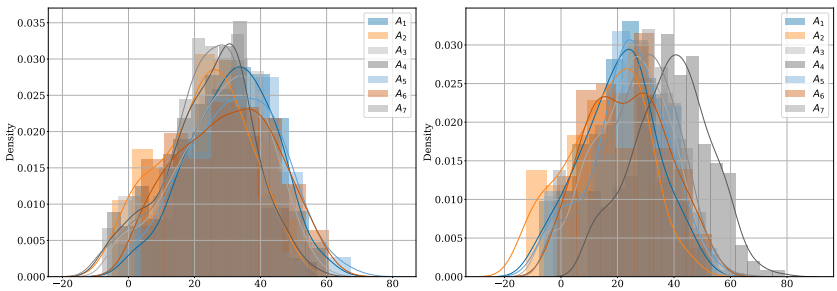


Figure 5: Distributions of treatment outcomes under two different latent states showing that the outcomes are approximately gaussian

B.3 Correctness Results

Shown in Figure 6 below.

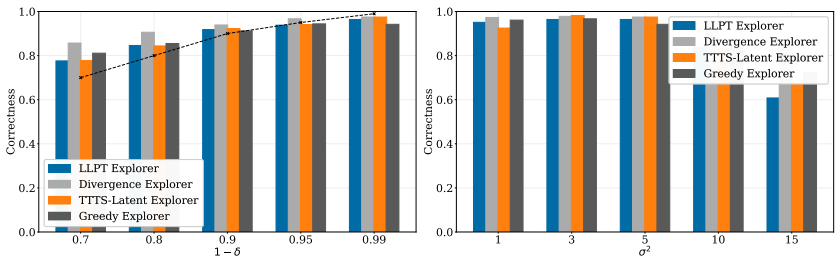


Figure 6: Correctness levels under (a) Varying δ levels; Dotted line marks the desired correctness level (b) Varying σ levels with $\delta = 0.01$

Thank you.