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

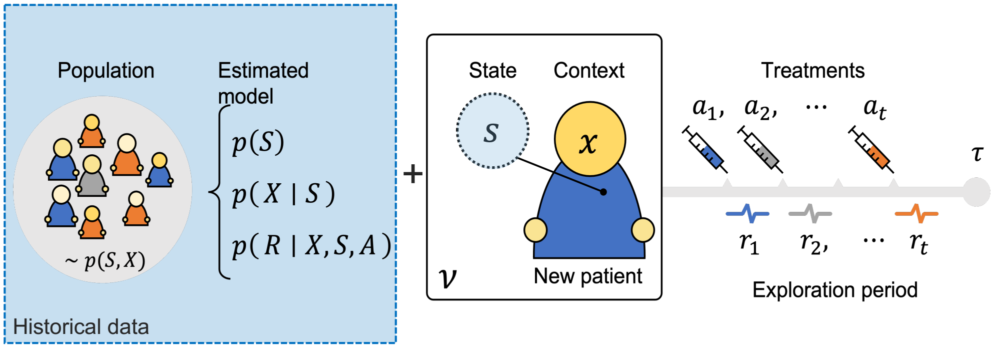


Figure 1: Illustration of the pure-exploration latent bandit problem and the example of treatment personalization. A population of patients have been observed in historical data to learn the distribution of latent states $P(S)$, $P(X|S)$ and the conditional reward the distribution $P(R|X, S, A)$. A new 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*.*

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

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, $\forall x, s : P(\mu_{\hat{a}_\tau,x,s} < \mu^*_{x,s} \mid X = x, S = s) \le \delta$, 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 $\text{KL}(\mu_{a,x,s} \,\|\, \mu_{a,x,s'}) = \text{KL}(p(R \mid a, x, s) \,\|\, p(R \mid a, x, s'))$ 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

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

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 x,

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

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

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

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,

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

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

Because $\mathrm{KL}(\mu_{a,x,s_t} \parallel \mu_{a,x,s'_t})$ 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, $\mathrm{Alt}_x(s) = \mathcal{S} \setminus \{s\}$. 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 $\mathrm{Alt}_x(s)$ is smaller compared to the case with no structure in, for example, Garivier and Kaufmann (2016). In our case, $\mathrm{Alt}_x(s)$ 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.

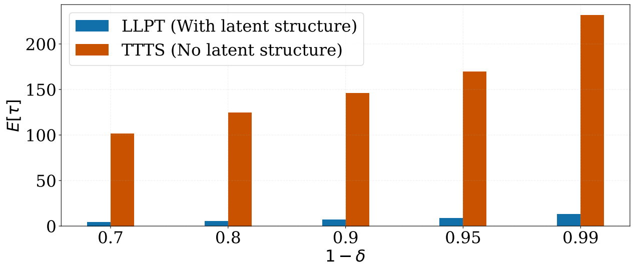


Figure 2: Using latent state structural information significantly reduces the expected number of trials $\mathbb{E}[\tau]$ required to identify an optimal treatment with confidence at least $1-\delta$ in a simulator of Alzheimer's disease progression.

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.

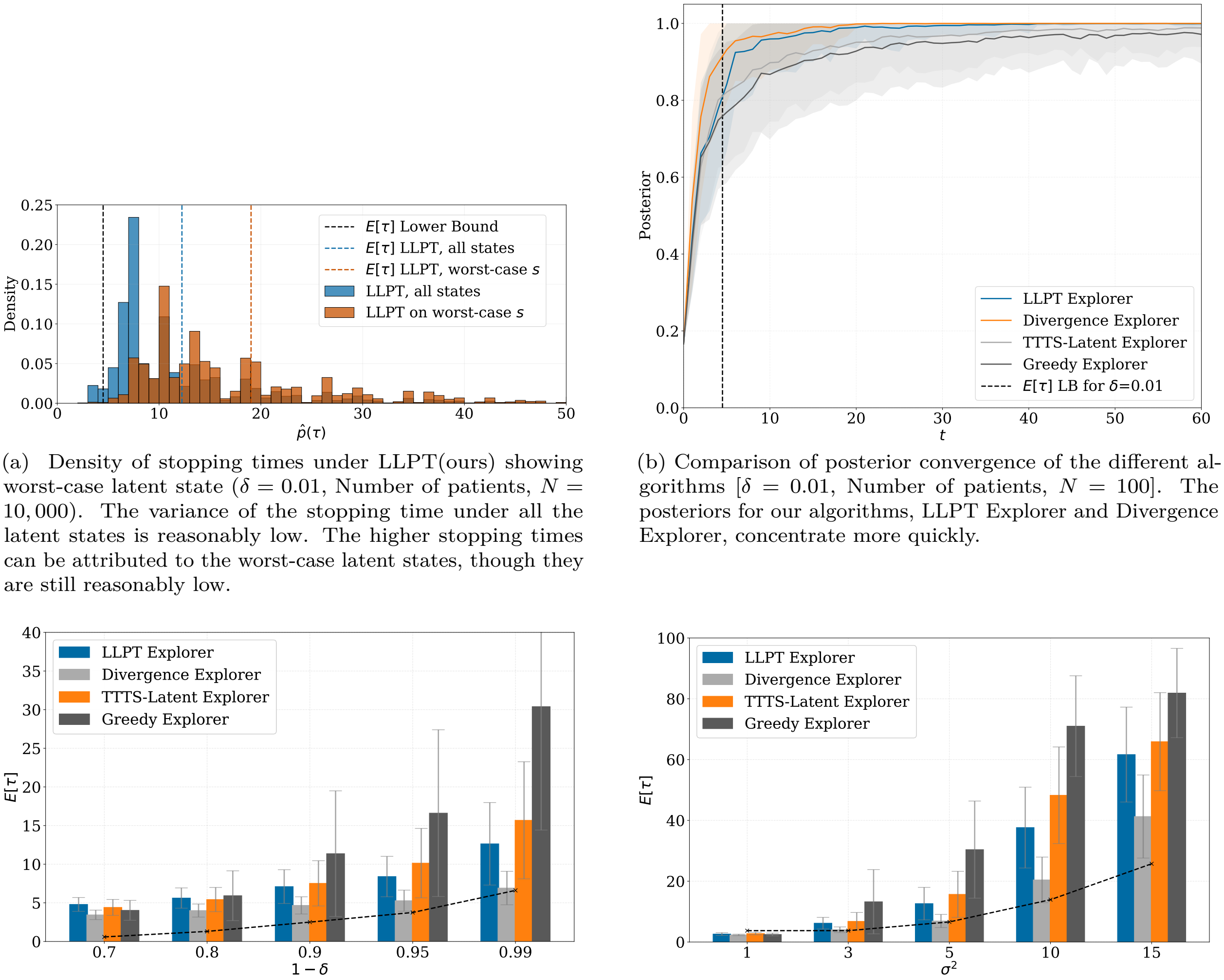
In Figure 3c, we study the average stopping time, $\hat{\mathbb{E}}[\tau]$ for all algorithms with access to the same latent variable model, under changing certainty level $1-\delta$. LLPT Explorer and Divergence Explorer are consistently more efficient than baselines, demonstrating benefit of the insights 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

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

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

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

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 $\epsilon_1 > 0$, $(\sum_{i=1}^t [z_i(s,s') - \mathbb{KL}^{R,a_i,x}_{s,s'}] + \epsilon_1)$ 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,

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

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

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

and, as a result,

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

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,

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

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', $\mathbb{KL}^{R,w^\zeta,x}_{s,s'} \geq C^\zeta(s,x)$.

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

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

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

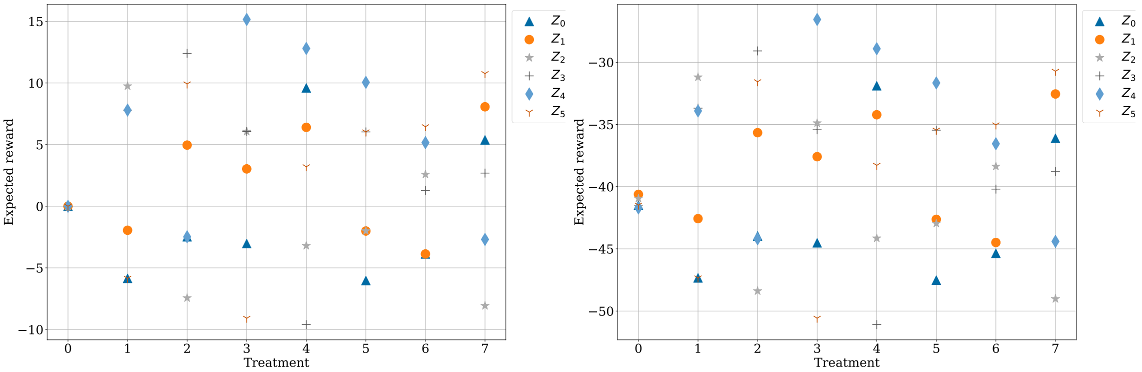


Figure 4: Structure of the means $\mu_{s,a}$ under different latent states. (a) Non-contextual rewards and (b) Contextual rewards

B.2 Outcome Distribution

Shown in Figure 5 below.

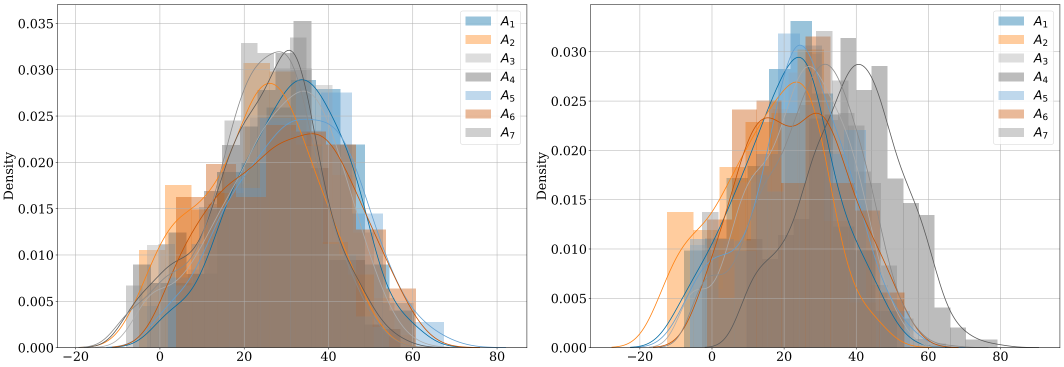


Figure 5: Distributions of treatment outcomes under two different latent states showing that the outcomes are approximately gaussian

B.3 Correctness Results

Shown in Figure 6 below.

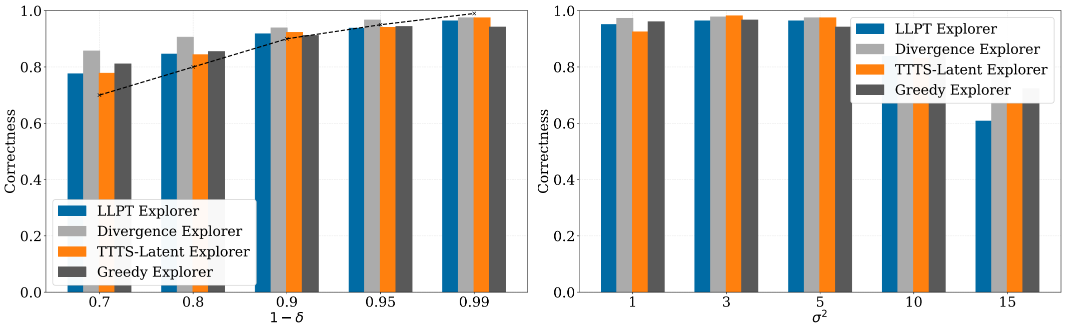


Figure 6: Correctness levels under (a) Varying δ levels; Dotted line marks the desired correctness level (b) Varying σ levels with $\delta = 0.01$