Linearizing Contextual Multi-Armed Bandit Problems with Latent Dynamics

by

Elliot Nelson

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2022

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Elliot Nelson is the sole author of material in this thesis, with the following exception: Debarun Bhattacharjya contributed to the generation of the data labelled as "umTS" in Figure 5.1, and wrote part of the text in the last paragraph on page 2 and the first paragraph on page 3.

Abstract

In many real-world applications of multi-armed bandit problems, both rewards and observed contexts are often influenced by confounding latent variables which evolve stochastically over time. While the observed contexts and rewards are nonlinearly related, prior knowledge of latent graphical structure can be leveraged to reduce the problem to the linear bandit setting. We develop a linearized latent Thompson sampling algorithm (L^2TS), which exploits prior knowledge of the dependence of observed contexts on the hidden state to build a least-squares estimator of the latent transition matrix, and uses the resulting approximate posterior beliefs over the latent space as context features in a linear bandit problem. We upper bound the error in reward parameter estimates in our method, demonstrating the role of the latent dynamics and evolution of posterior beliefs. We also demonstrate through experiments the superiority of our approach over related bandit algorithms. Lastly, we derive a theoretical bound which demonstrates the influence of the latent dynamics and information theoretic structure of the problem on Bayesian inference over the latent space. Overall, our approach uses prior knowledge to reduce a complex decision-making problem to a simpler problem for which existing solutions and methods can be applied.

Acknowledgements

I am grateful to my collaborators Tian Gao, Miao Liu, and Djallel Bouneffouf for their support and insight, and in particular to Debarun Bhattacharjya for his support in designing experiments and in preparing the work in this thesis for publication.

I am especially grateful to Pascal Poupart for his patient guidance of my research throughout the course of my Master's degree, and for his thoughtful attention to my half-baked ideas.

Most of all, I am grateful to my wife Kelly for her patience with me during times of stress and uncertainty.

Table of Contents

List of Figures

Chapter 1

Introduction

Many real-world sequential decision-making problems involve the following two challenges:

- *Partial observability.* Real-world data are in general influenced by unobserved latent variables. Often, latent variables correspond to higher levels of abstraction which must be modeled (at least implicitly) in order to generate accurate predictions of future observations.
- *Non-stationarity.* Real-world sequential data are in general not independent and identically distributed (i.i.d.). For time-series data, the i.i.d. assumption is generally violated due to underlying temporal dynamics which introduce correlations between data at different times as well as non-stationarity in the distribution of observed data over time. Often, this is due to dynamical evolution of unobserved latent variables.

Sequential decision-making methods which effectively optimize a target performance metric must model these components of the data-generating process in order to more accurately predict the consequences of decisions.

In this thesis, we address these challenges in the multi-armed bandit setting. Multi-armed bandits have been successfully applied in domains such as healthcare [22, 65], finance [53], and recommender systems [64]. While limited by the constraint that actions do not influence any underlying state variables, the bandit setting allows for more tractable theoretical analysis of algorithms (especially when dynamics and partial observability already lead to a nontrivial level of complexity), and is a stepping stone towards understanding of reinforcement learning [55] algorithms in the more general setting of partially observable Markov decision processes (POMDPs) [54].

We focus on a contextual multi-armed bandit problem with the graphical structure shown in Figure 3.1 (see Chapter 3 below). In this problem, an unobserved latent state (denoted z_t) evolves over time according to an (unknown) state transition model. At each timestep t, a context (denoted x_t) is generated conditional on the current latent state, the agent selects an action a_t after observing x_t, and a reward r_t is generated conditional on the agent's action and the current latent state. This problem may be viewed as a subset of partially observable Markov decision processes (POMDPs), to which the graphical model can be generalized by adding an additional edge from the agent's action a_t to the next state z_{t+1}, allowing control of the latent state. In this light, our problem is the mildest possible simplification of a POMDP to the bandit framework (in which the agent cannot influence state transitions), and as such, is an important bridge problem connecting the space of bandit decision problems and algorithms to more general decision-making problems. It is also worth noting that this problem setting is a slight generalization of hidden Markov models, which can be obtained by reducing the action space to a single action, and considering contexts and rewards to constitute a joint observation variable which is emitted by a hidden Markov state.

In this problem, the observed context is correlated with subsequent rewards via the latent state (conditional on which they are independent), and thus, selecting actions which maximize the expected reward generically requires inference of the current latent state. Our approach to the problem will combine a particular method of inference with existing multi-armed bandit methods for learning how the latent variable influences rewards.

Our setting applies generally to situations where explicitly modeling a latent variable is natural and beneficial, possibly due to a causal mechanism where the context is 'caused' by the latent variable. This often implies that the observed context contains useful information about the underlying latent variable, beyond what one could infer from rewards alone, making it possible to better predict rewards by inferring latent states from observations. Consider the following illustrative real-world applications:

- An interactive AI agent for personalized education chooses material to help a student's evolving state of knowledge, using observations such as the time taken to answer questions.
- A rover on a mission explores blocks of land, taking samples that provide information about the ore grade and choosing mining strategies for each block in real time.
- A recommender system selects items for users with evolving latent preferences or values, potentially using observable signals such as behavior patterns.

In cases such as these, the latent graphical structure of the problem is often known in advance. Modeling the causal mechanism enables one to use domain expertise or pre-existing data to estimate the probability distribution of the context conditioned on the latent variable, $p(x|z)$, in order to accelerate inference of the latent state. For instance, a medical professional may provide estimates of likelihoods of a patient's underlying disease condition z generating different possible observed symptoms x. Alternatively, in settings where high-dimensional context data is already available, estimates for $p(x|z)$ can be pre-trained offline and transferred to novel online settings in which latent dynamics or rewards may be altered but the same causal mechanism $p(x|z)$ is at play. (Another potential way to obtain this knowledge is by using data when latent variables can be inferred after a delay, e.g. retrospective analysis of latent financial conditions.) At the same time, while the particular algorithm which we introduce for this problem makes use of assumed prior knowledge of this kind in order to accelerate learning, our high-level approach is more general, and can be extended straightforwardly to settings where less prior knowledge is available.

1.1 Overview of Thesis Contributions

In Chapter 2, we describe the multi-armed bandit framework, review relevant literature, including the linear bandit setting and linear Thompson sampling algorithm of Agrawal and Goyal [4], upon which we will build.

In Chapter 3, we introduce our problem setting, a contextual multi-armed bandit problem involving a dynamical latent state which influences reward variables and additional context observations, show that prior knowledge of the graphical structure of the problem (along with partial knowledge of environment parameters) can be used to reduce the problem to the linear bandit setting, and derive an upper bound on the error of a linear regression estimator for unknown reward parameters, constructed from posterior probabilities for current and past latent states.

In Chapter 4 we build on the results of Chapter 3 and develop a novel Thompson sampling algorithm, Latent Linear Thompson Sampling (L^2TS), which performs Bayesian inference over the unobserved latent state and uses the reward estimator introduced in Chapter 3 as part of the linear bandit algorithm of Agrawal and Goyal [4]. We present results from experiments in low-dimensional synthetic environments which show that our algorithm is able to outperform baseline algorithms designed for contextual multi-armed bandit algorithms with non-stationarity.

In Chapter 5, we study the influence of the conditional distributions $p(x|z)$ on inference of the latent state. We derive an upper bound on regret relative to an oracle algorithm which observes the latent state, for a related Thompson sampling algorithm, which captures the effects of dynamical and information theoretic structure of the environment on task difficulty.

We conclude in Chapter 6, summarizing our contributions and outlining several directions for future research.

Chapter 2

Background

In this chapter, we first describe in Section 2.1 the basic setup for multi-armed bandit problems, and subsequently (in Sections 2.2 and 2.3) review existing work in two classes of multi-armed bandit problems (latent and non-stationary bandits) which extend this basic setup by modifying the graphical structure and time-dependence of the data generating process, respectively. (In Chapter 3 below, we will define our problem setting at the intersection of these two problem settings.) Lastly, in Section 2.4, we introduce the linear bandit problem setting and review a particular algorithm for this setting which we will adapt for the more general setting of latent and non-stationary multi-armed bandits.

2.1 Multi-Armed Bandits

In multi-armed bandit (MAB) problems [36, 7, 38], an agent makes a sequence of decisions with the goal of optimizing a reward signal. At each timestep $t = 1, 2, ...$, the agent selects an action $a_t \in \{1, ..., K\}$ among K possible discrete-valued actions (bandit arms), and subsequently receives a reward $r_t \in \mathbb{R}$ generated from an action-dependent distribution, $r_t \sim p(r|a = a_t)$. The goal of the agent is to maximize its return, the sum of rewards $\sum_{t=1}^{T} r_t$, up to a time horizon T. We will focus on the limit of an infinitely long time horizon, $T \to \infty$.

In the more general class of *contextual* multi-armed bandit problems, the agent receives an observation or context x_t at each round, prior to selecting an action, and the subsequent reward is generated from a context-dependent distribution, $r \sim p(r|x = x_t, a = a_t)$.

Any particular bandit algorithm maintains a policy in the form of a probability distribution for selecting actions, depending on rewards and contexts observed thus far. Defining

$$H_t := (x_1, a_1, r_1, ..., x_{t-1}, a_{t-1}, r_{t-1}, x_t) \tag{2.1}$$

as the history of observations, actions, and rewards up to time t (including the most recent context x_t), actions are sampled from a distribution,

$$a_t \sim \pi(\cdot|H_t). \tag{2.2}$$

The *optimal policy* $\pi^\star$ is commonly defined as the policy which selects the action with highest expected reward,

$$\pi^\star(a|H_t) = \mathbf{1}(a = \arg\max_{a'} \mathbb{E}[r_t|a_t = a'; H_t]), \tag{2.3}$$

where the expectation is taken with respect to the true reward distribution.

The performance of bandit algorithms is typically described in terms of *regret*, the difference in returns earned by a given policy π relative to the optimal policy up to time T,

$$\mathcal{R}_\pi(T) := \sum_{t \leq T} \left(\mathbb{E}_{\pi^\star}[r_t] - \mathbb{E}_\pi[r_t]\right). \tag{2.4}$$

Unless otherwise noted, when referring to regret we will always have in mind the cumulative expected regret as defined in Eq. (2.4).

Regret bounds and problem dependence. Theoretical support for bandit algorithms is often sought in the form of upper bounds on the regret $\mathcal{R}_\pi(T)$ incurred by the algorithm within time horizon T, or, relatedly, on the expected number of suboptimal actions selected. The bulk of theoretical upper bounds on expected regret are *problem-independent*, meaning that they do not depend on the particular reward distributions $p(r|a)$ – or $p(r|x, a)$ in contextual settings – of a particular problem instance, aside from constraints (e.g. on the tails or domain of support of distributions) which apply to a wide range of reward distributions. In contrast, *problem-dependent* or instance-dependent bounds leverage structure or features of the reward distribution of a particular problem which influence the problem difficulty or performance of a given algorithm. In particular, *gap-dependent* bounds make use of the differences (gaps) in expected rewards between optimal and suboptimal actions, $\Delta_a = \mathbb{E}[r|a^\star] - \mathbb{E}[r|a]$ where $a^\star = \arg\max_a \mathbb{E}[r|a]$ in the non-contextual case. In the contextual case, (i) the gaps become functions of the context, and (ii) the structure of the context distribution as well as reward distributions can affect problem difficulty, leading to additional dependencies in problem-dependent bounds.

Exploration and information gain. A key challenge faced by multi-armed bandit algorithms (and sequential decision making algorithms in general) is that of *exploration*. Suboptimal actions must be explored enough times to be ruled out, but should not be chosen more than necessary. Selecting suboptimal actions will, in expectation, yield valuable information about the corresponding reward distributions which will eventually be used to rule those actions out with higher confidence. The challenge of exploration is the challenge of trading off this information gain (and resulting long term expected rewards) with the cost of lower expected rewards in the short term.

2.1.1 Bandit Algorithms

Two of the most common approaches to exploration are upper confidence bound methods and Thompson sampling.

Upper confidence bound (UCB) approaches maintain estimates $\hat{\mu}(a)$ of the expected reward for each action a, along with corresponding estimates $\hat{u}(a) \geq 0$ of uncertainty of the expected reward, and select at each timestep the action which maximizes $\hat{\mu}(a) + \hat{u}(a)$, acting optimistically with respect to more uncertain actions.

Thompson sampling (TS) [56, 17, 48], the approach we will follow in this thesis, is a Bayesian approach to exploration which samples actions according to the posterior probability of their being optimal. Thompson sampling maintains a posterior distribution over possible environments (i.e. reward distributions, or a subset of parameters or features thereof). At each timestep, Thompson sampling (i) samples the current posterior distribution over possible models of the environment, (ii) selects the action a_t which yields highest expected reward under the sampled model, and (iii) performs a Bayes update on the posterior using the reward r_t received from action a_t.

In this thesis, we focus on Thompson sampling instead of UCB-type algorithms. In particular, we build on the linear Thompson sampling algorithm of Agrawal and Goyal [4, 3], discussed in greater detail in Section 2.5.

2.2 Latent Bandits

The present work establishes a connection between linear bandit problems and *latent bandit* problems, in which rewards are influenced by one or more latent variables. In contextual latent bandit settings, latent variables may or may not also influence observed contexts.

In our particular latent bandit setting, described in Section 3.1, both rewards and observed contexts are influenced by a discrete, categorical latent variable, which introduces correlations between rewards and contexts.

Here, we note two recent works which study related but distinct latent bandit settings. Maillard and Mannor [44] consider a specific case of our setting in which a categorical latent variable indexes different clusters, observed contexts are discrete types, and the set of types is partitioned into disjoint subsets corresponding to distinct clusters. In comparison, we consider more general latent-conditioned context distributions (for which observed contexts may be generated from various latent states), focusing on the case where prior knowledge of these distributions is available. Zhou and Brunskill [63] consider a related setting in which rewards are linearly related to observed contexts (as in Section 2.4), with the parameters of the linear relationship influenced by a latent state. In their setting, the reward depends directly on the observed context, as well as the latent state. In contrast, in our setting (which may be viewed as a subset of partially observable Markov decision processes, or POMDPs) the reward depends only on the latent state, and the information value of observed contexts is only for inference of the latent state. While our setting also reduces to a linear bandit problem under certain conditions, the relationship between rewards and observed contexts can remain arbitrarily nonlinear.

Also worth noting is Hong et al. [31], which studies a related but non-contextual latent bandit problem, and (like the present work) applies Thompson sampling. In Section 2.3 below, we discuss the relationship of our work to the non-stationary extension [32] of [31].

2.2.1 Recommender Systems

The graphical structure of our problem, with a latent confounder influencing observation and reward, is shared in some literature of bandit algorithms for recommender systems. In particular, [50] obtain a regret bound for an ϵ-greedy algorithm which leverages the low dimensionality of the latent space. Relatedly, [35] obtain a regret bound for a Thompson sampling algorithm in a similar setting with unknown model parameters and Gaussian reward likelihoods. In comparison, our work considers more general likelihoods which are partially known in advance, and generalizes to non-stationarity in the latent state, which introduces a complexity which precludes straightforward extension of regret bounds from the i.i.d. (stationary) setting. At the same time, our analyses in the non-stationary setting demonstrate that knowledge of a low-dimensional latent space can be leveraged for improved performance, similar to results for recommender systems in i.i.d. settings.

2.3 Non-stationary Bandits

The problem we introduce in Section 3.1 is also a *non-stationary* multi-armed bandit problem. In stationary settings, which we've implicitly assumed above, the reward distribution $p(r|x, a)$ is constant in time, and does not depend on contexts, rewards, or actions prior to the current round (timestep). That is, when conditioning on a particular action, rewards are independent and identically distributed (i.i.d.) across time. In non-stationary settings, on the other hand, the reward distribution may become time-dependent, and may also introduce correlations between rewards (or contexts) at different times.

In the non-stationary latent bandit setting we introduce in Section 3.1, non-stationarity arises from Markovian transitions in a discrete latent state. We will occasionally refer to this case as dynamical, a term which we use interchangeably with non-stationarity. (We will assume the transition probabilities of the latent state, along with other parameters, to be constant in time, and thus use "non-stationary" in reference to the latent state itself, rather than to its transition parameters.)

In piecewise stationary bandits problems [29, 24, 59], reward distributions for one or more arms change in a discrete, discontinuous manner at certain timesteps, but remain fixed between these change points. Our latent bandit problem may be viewed as a piecewise stationary bandit problem in which change points correspond to state changes of an underlying latent state. As such, we assume that additional structure (a fixed number of latent states z, and corresponding context distributions $p(x|z)$) is known by the agent, allowing for faster detection of change points relative to other piecewise stationary bandit methods.

Hong et al. (2020).

The recent work which is most closely related to this thesis is that of Hong et al. [32] (see also [31]), who study Thompson sampling for a related non-stationary latent bandit problem. In their case, an approximate posterior over discrete latent state histories and unknown parameters is updated using rewards rather than context observations, which are assumed unaffected by the latent state. The graphical model of our setting differs slightly, in that the latent state influences context data as well as rewards. More significantly, the method of [32] maintains a particle-based posterior over possible latent transition matrices and context distributions, whereas our approach uses point estimates of these parameters. The former approach naturally has an advantage when uncertainty in the latent transition function can be leveraged to explore actions more effectively. However, maintaining a good approximate posterior over both latent histories and parameters is challenging. In particular, we find that the particle filtering method of [32], which updates particle weights but

not particle positions, does not perform well empirically. Our approach is more computationally lightweight, algorithmically simpler, and makes use of a task-relevant (albeit more limited) measure of uncertainty in the form of marginal posterior beliefs over the current latent state and reward parameters (see Section 6.1 for more discussion).

In addition to algorithmic work, Hong et al. [32] derive problem-independent regret bounds for their Thompson sampling algorithms. These bounds assume sub-Gaussian rewards but do not otherwise make use of the structure of the reward distribution or other problem parameters. The resulting bounds on cumulative regret are sublinear at small t before the latent state evolves significantly, but scale as $t^{7/6} \log t$ as $t \to \infty$ (a consequence of the assumption that the number of latent state changes grows linearly with t). In comparison, in chapter 5 we derive a problem-dependent bound for an equivalent definition of regret, which leverages the information theoretic structure of the conditional distributions $p(x|z)$, and scales linearly with t (an inevitable consequence of the definition of regret with respect to an oracle policy which access the true latent state), an improvement over the asymptotic behavior of the bounds in [32]. Empirically, we found (see Chapter 4) that the algorithm of [32] struggles in practice, likely because it uses a naive particle filtering method which only updates weights assigned to a fixed, non-adaptive set of possible parameter vectors.

Lastly, we also note another recent work of Hong et al. [30], which considers a slightly generalized version of our non-stationary latent bandit setting, with an additional direct influence of observed contexts on rewards, and an unknown number of latent states. In contrast to our approach, [30] focuses on off-policy learning, and assumes that context distribution parameters and latent transition probabilities are estimated offline. While our setting is slightly less general, our focus on online learning is complementary to [30].

2.4 Linear Bandits

Linear stochastic bandit problems form a much-studied subset of contextual MAB problems. In these cases, an observed context feature vector is assumed to be linearly related to the expected reward (conditional on the context). We focus on the linear stochastic K-armed bandit problem, in which K discrete actions are available, in contrast to more general problems in which continuous-valued actions are available. In Chapters 3 and 4 below, we will identify a connection between linear bandit problems and the latent non-stationary bandit setting introduced in Section 3.1 below, and will apply algorithmic methods and theoretical analysis of the former to the latter.

We will work with a multi-armed linear bandit setting which is a slight modification of the typical setting in the literature [4], as follows. At each timestep t, a context feature vector $c_t \in \mathcal{C} \subset \mathbb{R}^d$ is observed, an action $a_t = a \in \{1, ..., K\}$ is selected, and a reward

$$r_t = c_t^\top \mu_\star^{(a)} + c_t^\top \eta_t^{(a)} \tag{2.5}$$

is observed. The reward is generated from a fixed distribution $p(r|c, a)$ with a mean value

$$\mathbb{E}_{r \sim p(\cdot|c,a)}[r] = c^\top \mu_\star^{(a)} \tag{2.6}$$

that is linear in both the context c and the unknown, action-dependent parameters $\mu_\star^{(a)} \in \mathbb{R}^d$. (Note that in other variations of the linear bandit setting, the same parameters μ may be shared across actions, while a separate per-action context $c_t^{(a)}$ may be observed.) The random noise vector $\eta_t^{(a)} \in \mathbb{R}^d$ has mean zero by definition, $\mathbb{E}[\eta_t^{(a)}] = 0$, and is assumed to be generated independently at each timestep by a fixed, time-independent distribution. However, the noise term $c_t^\top \eta_t^{(a)}$ in Eq. (2.5) is assumed to depend linearly on the context c_t, which is time-dependent and may be correlated across time.

In order to maximize returns, the agent must use the sequential context data $c_{1:t}$ to learn the unknown mean reward parameters $\mu_\star^{(a)} \in \mathbb{R}^d$ for each action a.[1] Given a context c, the corresponding optimal action is

$$a^\star(c) := \arg\max_a \mathbb{E}[r|c, a] = \arg\max_a (c^\top \mu_\star^{(a)}). \tag{2.7}$$

We will refer to the optimal action at timestep t as $a_t^\star := a^\star(c_t)$.

2.5 Thompson Sampling

In the linear bandit setting, Thompson sampling methods maintain and sample a posterior over the unknown parameters $\mu^\star := \{\mu_\star^{(a)}\}_{a=1}^K$. In Chapter 4, we will introduce a novel algorithm which uses the linear Thompson sampling (LinTS) algorithm of [4] as a subroutine. The LinTS algorithm, reproduced here as Algorithm 1, maintains for each action a multivariate normal posterior with mean $\hat{\mu}^{(a)}$ and covariance $(B_\mu^{(a)})^{-1}$ (defined below). After observing a context vector c_t at timestep t, LinTS samples reward parameters

$$\mu^{(a)} \sim \mathcal{N}(\hat{\mu}^{(a)}, (B_\mu^{(a)})^{-1}) \tag{2.8}$$

[1] In other variations of the linear bandit setting, the same parameters μ may be shared across actions, while a separate per-action context $c_t^{(a)}$ may be observed.

for each action a, and selects the corresponding optimal action, $a_t^\star = \arg\max_a c_t^\top \mu^{(a)}$.

The mean vector $\hat{\mu}^{(a)}$ of the posterior used for Thompson sampling in Eq. (2.8) at time T is the least-squares estimator,

$$\hat{\mu}^{(a)} := (B_\mu^{(a)})^{-1} f_\mu^{(a)} \tag{2.9}$$

where

$$f_\mu^{(a)} := \sum_{t=1}^{T} \mathbf{1}(a_t = a) c_t r_t, \tag{2.10}$$

$$B_\mu^{(a)} := \lambda_\mu \mathbb{1} + \sum_{t=1}^{T} \mathbf{1}(a_t = a) c_t c_t^\top, \tag{2.11}$$

where the hyperparameter $\lambda_\mu \geq 0$ can be used to ensure that $B_\mu^{(a)}$ is positive definite at small T. (We will often suppress the implicit T-dependence of these quantities.)

Eqs. (2.10)-(2.11) (and Eq. (2.9)) can equivalently be written recursively. After observing context c_t, selecting action a_t, and receiving reward r_t, the mean and covariance for $a = a_t$ are updated,

$$B_\mu^{(a)} \leftarrow B_\mu^{(a)} + c_t c_t^\top \tag{2.12}$$

$$f_\mu^{(a)} \leftarrow f_\mu^{(a)} + c_t r_t \tag{2.13}$$

$$\hat{\mu}^{(a)} \leftarrow (B_\mu^{(a)})^{-1} f_\mu^{(a)} \tag{2.14}$$

This update rule can be recovered as a Bayes update under the assumption of a Gaussian reward likelihood,

$$P(r|c, a; \mu) \propto \exp\left[-(r - c^\top \mu^{(a)})^2 / 2(\tilde{\sigma}_r^{(a)})^2\right].$$

Given the Gaussian prior distribution,

$$P(\mu^{(a)}) \propto \exp\left[-\frac{1}{2(\tilde{\sigma}_r^{(a)})^2}(\mu^{(a)} - \hat{\mu}^{(a)})^\top B_\mu^{(a)} (\mu^{(a)} - \hat{\mu}^{(a)})\right], \tag{2.15}$$

and noting that

$$(\mu - \hat{\mu})^\top B_\mu (\mu - \hat{\mu}) = \mu^\top B_\mu \mu - \mu^\top f_\mu - f_\mu^\top \mu + const.,$$

where we've suppressed action-dependence for simplicity, we see that making the update in Eqs. (2.12)-(2.14) is equivalent to multiplication by the likelihood, Eq. (2.15), for observing

reward $r = r_t$ conditional on $c = c_t$ and action a, up to μ-independent normalization terms in the exponent.

Algorithm 1: Linear Thompson Sampling

Input:

$\lambda_\mu > 0$, $\tilde{\sigma}_r^{(a)} > 0$ for $a \in \mathcal{A}$

$\hat{\mu}^{(a)} = \mathbf{0}_d$, $f_\mu^{(a)} = \mathbf{0}_d$, $B_\mu^{(a)} = \lambda_\mu \mathbb{1}_d$, for $a \in \mathcal{A}$

for $t \leftarrow 1, 2, \ldots$ **do**

Observe context c_t

Sample $\mu^{(a)} \sim \mathcal{N}(\hat{\mu}^{(a)}, (\tilde{\sigma}_r^{(a)})^2 (B_\mu^{(a)})^{-1})$ for $a \in \mathcal{A}$

Select action $a = \arg\max_{a'} c_t^\top \mu^{(a')}$

Observe reward r_t

Update mean reward estimates:

$B_\mu^{(a)} \leftarrow B_\mu^{(a)} + c_t c_t^\top, \quad f_\mu^{(a)} \leftarrow f_\mu^{(a)} + c_t r_t$

$\hat{\mu}^{(a)} = (B_\mu^{(a)})^{-1} f_\mu^{(a)}$

Chapter 3

Linearizing Latent Bandits

In this chapter, we describe our problem setting with a dynamical latent state which influences context observations and rewards, identify conditions under which this problem setting can be partially reduced to the linear bandit setting of Section 2.4, and derive a high-probability bound, Theorem 1, on the error in linear regression reward estimation when applied in the latent bandit setting.

3.1 Problem Setting

We consider the non-stationary bandit environment of Figure 3.1 in which a dynamical latent state z acts as a confounder of observations (or contexts) x and rewards r. The figure is represented as an influence diagram, which is a graphical model for decision making under uncertainty [33]. At any epoch, context x is observed before selecting action a, and reward r depends on a and z.

While the context and reward may be either discrete or real-valued[1], the latent state $z \in \mathcal{Z} = \{1, ..., Z\}$ and action $a \in \mathcal{A} = \{1, ..., K\}$ are assumed to be discrete. The latent state z evolves stochastically according to a transition matrix $\Phi^\star$ (assumed to be ergodic) with elements, $p(z_t = z' | z_{t-1} = z; \phi^\star) = \phi^\star_{z,z'}$. The equilibrium distribution $\rho^{(\phi)}_{\text{eq}}(z)$ is the stationary distribution, $\Phi \rho^{(\phi)}_{\text{eq}} = \rho^{(\phi)}_{\text{eq}}$. (For any categorical distribution $p(z)$, we will denote by $p \in \mathbb{R}^Z$ the vector whose elements are the probabilities $p(z)$.) Given z, an observed

[1] We denote context as a scalar for simplicity but our work is equally applicable to settings with high-dimensional observations.

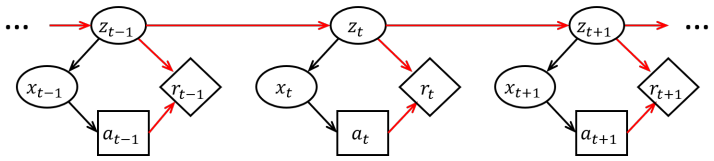

Figure 3.1: An influence diagram representation of the non-stationary version of our latent bandit setting. The latent state z changes dynamically while context x is observed at the time of choosing action a (rectangle), represented by the informational arc from x to a. Reward r (diamond) is a function of a and z. Black and red conditional edges denote known and unknown (or learned) conditional distributions respectively.

context x is generated from a conditional distribution $p(x|z;\theta^\star)$ with parameters $\theta^\star$. We assume access to a good but possibly imperfect approximation θ to $\theta^\star$, which may be available via domain expertise, or via offline samples $x_i \sim p(x|z)$ and accompanying labels z of the generating distribution. The reward-generating conditional distributions $p(r|z,a)$, on the other hand, are unknown. We define the expected reward conditional on latent state z as

$$(\mu_\star^{(a)})_z := \mathbb{E}_{r\sim p(\cdot|z,a)}[r], \tag{3.1}$$

and use $\mu_\star^{(a)}$ to denote the action-wise latent-space vectors of mean rewards. We will denote the variance of $r \sim p(\cdot|z,a)$ as $\mathrm{Var}[r|z,a]$.

Our algorithmic methods and theoretical analysis make use of the information theoretic structure of the context distributions $p(x|z)$. In particular, in Section 4.1 we will use the cross-entropies of the context distributions,

$$H_{z^\star,z}(\theta) := \mathbb{E}_{x\sim p(\cdot|z^\star;\theta)}\left[-\log p(x|z;\theta)\right] \tag{3.2}$$

to construct a model estimate of the transition matrix given prior knowledge of $\theta^\star$.

3.1.1 Bayesian Inference of the Latent State

Our algorithm relies on the approximation and use of a posterior belief, $p_t(z|x_{1:t}) := p(z_t = z|x_{1:t})$ over the current latent state, which is a categorical distribution represented as a Z-dimensional vector. Given transition model $p(z'|z;\hat{\phi})$ and observation model $p(x|z;\theta)$, it can be updated every epoch:

$$\hat{p}_t(z|x_{1:t}) \propto \sum_{z'} \hat{p}_{t-1}(z'|x_{1:t-1})\hat{\phi}_{z,z'}p(x_t|z;\theta) \tag{3.3}$$

where $\hat{p}, \hat{\phi}$ denote model estimates. We will distinguish the model posterior $\hat{p}$ from the "true" posterior

$$p_t^\star(z) := p(z_t = z | x_{1:t}; \phi^\star, \theta^\star, \rho_0), \tag{3.4}$$

which uses ground truth parameters and the true prior, $p_0^\star(z) := \rho_0(z)$.

The optimal policy $\pi^\star$ is then defined as the policy which selects, at every timestep, the action with highest expected reward, given the true parameters, $a_t^\star := \arg\max_a (p_t^\star)^\top \mu_\star^{(a)}$. We will quantify performance with expected cumulative regret, defined – for any policy π – as the loss in expected rewards after T timesteps relative to the optimal policy: $\mathcal{R}_\pi(T) := \sum_{t \leq T} (\mathbb{E}_{\pi^\star}[r_t] - \mathbb{E}_\pi[r_t])$.

3.2 Reduction to the Linear Bandit Setting

We now exploit the linear relationship between rewards and probabilities over the latent space to show that the latent bandit problem of Section 3.1 can be reduced to the linear bandit setting of Section 2.4. We show this in the case where the transition probabilities $p(z'|z;\phi)$ and context likelihoods $p(x|z;\theta)$ are known in advance, and comment below on the more general case where they are unknown.

Lemma 1. *When the true model parameters $(\theta^\star, \phi^\star)$ and initial latent state probabilities $\rho_0(z) = p(z_0 = z)$ in the model from Figure 3.1 are known, the latent bandit setting of Section 3.1 reduces to the linear bandit setting of Section 2.4.*

Proof. Conditional on a sequence of observations $x_{1:t}$ in the latent bandit setting and action $a_t = a$, the reward r_t is generated from the mixture distribution

$$p(r_t = r | a_t = a, x_{1:t}; \theta^\star, \phi^\star) = \sum_z (c_t)_z p(r|z,a), \tag{3.5}$$

where we have defined $c_t \in \mathbb{R}^Z$ as the vector with elements equal to the posterior probabilities

$$(c_t)_z := p(z_t = z | x_{1:t}; \theta^\star, \phi^\star). \tag{3.6}$$

A sample from this distribution can be obtained as a mixture of samples from $p(r|z,a)$,

$$r_t = \sum_z (c_t)_z \big((\mu_\star^{(a)})_z + (\eta_t^{(a)})_z \big) = c_t^\top (\mu_\star^{(a)} + \eta_t^{(a)}), \tag{3.7}$$

where $(\eta_t^{(a)})_z \sim p(r - (\mu_\star^{(a)})_z | z, a)$ has mean zero. Thus, the reward takes the form of Eq. (2.5), with $d = Z$ being the number of latent states, c_t defined in Eq. (3.6), $\mu_\star^{(a)} \in \mathbb{R}^Z$ being

the vector of mean rewards $(\mu_\star^{(a)})_z$, and $\eta_t^{(a)}$ being a vector of zero-mean noise generated from element-wise distributions $p(r - (\mu_\star^{(a)})_z | z, a)$. □

Lemma 1 shows that the posterior belief over the current latent state z_t can be viewed as a compression of the context history $x_{1:t}$ into a (nonlinearly) transformed context variable which is related linearly to rewards. Since Lemma 1 assumes access to the true parameters $(\theta^\star, \phi^\star)$, in general it will only apply in the asymptotic limit $(t \to \infty)$ in which $(\theta^\star, \phi^\star)$ have been learned. Prior to this asymptotic regime, error in model estimates of these parameters will corrupt the context features c_t in the corresponding linear bandit problem with noise and/or systematic bias.

In chapter 4, we will build on Lemma 1 to develop a latent bandit algorithm which estimates rewards at time T, Eq. (2.9), with contexts $c_t \to p_t^\star$ as in Eq. (3.6), that is,

$$\hat{\mu}^{(a)} = (B^{(a)})^{-1} f_\mu^{(a)}, \tag{3.8}$$

with

$$f_\mu^{(a)} = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p_t^\star r_t, \tag{3.9}$$

$$B^{(a)} = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p_t^\star (p_t^\star)^\top. \tag{3.10}$$

In sections 3.3-3.4 below, we state and derive a high-probability bound on the error in the least-squares estimator, Eq. (3.8), thus showing that linear bandit methods can be effectively applied in the latent bandit setting.

We end this section by noting that the space of context vectors c_t, or equivalently posterior beliefs $p_t^\star$ (see Eq. (3.6)), is partitioned into subspaces – denoted $\mathcal{P}_{a^\star}$ – for which action $a^\star$ is optimal, i.e. $a^\star = \arg\max_a c_t^\top \mu_\star^{(a)}$. This context-dependence of the optimal action will play a role in our derivation of Theorem 1 below.

3.3 Bound on Estimator Error

In this section, we demonstrate that linear bandit reward estimation can be effectively applied to the non-stationary latent bandit setting from Figure 3.1. Due to the time-dependence of the latent state, the reduction to the linear bandit setting described in

Section 3.2 results in contexts $c_{1:t}$ and rewards $r_{1:t}$ which are not i.i.d. across time. Here, we state a result which shows that reward estimation via reduction to the linear bandit setting, as defined in Eq. (3.8), will converge to the true reward parameters $\{\mu_\star^{(a)}\}$ given a sufficiently long time horizon:

Theorem 1. *Assuming that (i) the latent state Markov chain is ergodic and $z_1 \sim \rho_{\text{eq}}^{(\phi)}(\cdot)$, and when (ii) the true parameters $(\theta^\star, \phi^\star)$ and initial state distribution $\rho_{\text{eq}}^{(\phi)}$ are known and are used to compute $\hat{\mu}^{(a)}$, Eq. (3.8), the error in $\hat{\mu}^{(a)}$ at time $t = T$ for any algorithm which selects the optimal action given $x_{1:T}$ with probability at least $\pi_{\min}$, is upper bounded,*[2]

$$\begin{aligned} &|\hat{\mu}_z^{(a)} - (\mu_\star^{(a)})_z| \\ &< \frac{2Z^2}{\pi_{\min}^2 \lambda_{\min}^{(a)}} \sqrt{\frac{1}{\delta \cdot T}\left(\sigma_{\text{eq}}^2 + ||\mu_\star^{(a)}||_1^2 \frac{4}{\gamma_{\phi^\star}}\left(1 + \log c_{\phi^\star}\right)\right)} \end{aligned} \tag{3.11}$$

for any z with probability at least

$$1 - \delta - \frac{8Z^3}{\pi_{\min}^2 \lambda_{\min}^{(a)}} \frac{1}{T\gamma_{\phi^\star}} (c + \log\log(1/\rho_{\min})) \tag{3.12}$$

for any $\delta \in (0, 1)$. Here, $c \approx 6.8$, $c_{\phi^\star}$ is a $\Phi^\star$-dependent numerical constant (see Appendix 3.4.2), $\rho_{\min} := \min_z \rho_{\text{eq}}^{(\phi)}(z)$ is the equilibrium probability of the least probable latent state, $\sigma_{\text{eq}}^2 := \max_a \sum_z \rho_{\text{eq}}^{(\phi)}(z)\text{Var}[r|z, a]$ is a measure of reward noise when the latent state is in equilibrium, $\lambda_{\min}^{(a)} = \lambda_{\min}^{(a)}(T)$ is the minimal eigenvalue of the action-wise asymptotic expected inverse covariance matrix[3]

$$\Sigma^{(a)}(T) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{x_{1:t} \sim \rho_{\text{eq}}^{(\phi)}}[\mathbf{1}(p_t^\star \in \mathcal{P}_a) p_t^\star (p_t^\star)^\top], \tag{3.13}$$

averaged over histories generated from the equilibrium distribution, and

$$\gamma_\phi := \min_{z_1, z_2} \sum_z \min(\phi_{z,z_1}, \phi_{z,z_2})$$

is the minimal mixing rate of a transition matrix Φ [15].

[2] Here, $||\mu||_\ell$ denotes the ℓ-norm of a vector μ.

[3] Recall that $\mathbf{1}(p_t \in \mathcal{P}_a)$ is the binary truth value of the statement that $a = \arg\max_{a'} p_t^\top \mu_\star^{(a')}$ is the optimal action given the posterior belief p_t.

Proof (Outline). Section 3.4 has the complete proof. The derivation relies primarily on a KL divergence contraction theorem for stochastic Markov processes from [15] to show that posterior probabilities used to compute the estimators $\hat{\mu}^{(a)}$ are approximately uncorrelated, $\mathbb{E}[p_t^\star(z)p_{t'}^\star(z')] \approx \mathbb{E}[p_t^\star(z)]\,\mathbb{E}[p_{t'}^\star(z')]$, over time separations $|t-t'|$ greater than the minimal mixing time $1/\gamma_{\phi^\star}$. Thus, the quantities $f_\mu^{(a)}$ and $B^{(a)}$ in Eqs. (3.9)-(3.10) are sums of approximately independent random variables over blocks of at least $1/\gamma_{\phi^\star}$ timesteps. We quantify this with upper bounds on the variances $\mathrm{Var}[f_\mu^{(a)}]$ and $\mathrm{Var}[B^{(a)}]$ across context and reward histories, apply Chebyshev's inequality to obtain high-probability bounds on the deviation of $f_\mu^{(a)}$ and $B^{(a)}$ from their expected values at large T, and derive an eigenvalue bound for the inverse matrix $(B^{(a)})^{-1}$ in order to upper bound the deviation of the product $\hat{\mu}^{(a)} = (B^{(a)})^{-1} f_\mu^{(a)}$ from $\mu_\star^{(a)}$. □

Theorem 1 describes the effect of the latent dynamics and resulting posterior beliefs $p_t^\star$ on reward parameter estimation. At times T sufficiently large compared to the mixing time $t_{\phi^\star} := 1/\gamma_{\phi^\star}$, correlations between posterior beliefs (i.e. the dependent variables in linear regression estimation of $\mu_\star^{(a)}$) at different times are small, and reward data are close to i.i.d., allowing for a $1/\sqrt{T}$ error reduction. The dependence on $\lambda_{\min}^{(a)}$ in Eq. (3.11), which approaches a fixed asymptotic value in the $t \to \infty$ limit where posterior vectors p_t are generated from a fixed asymptotic distribution, captures the benefit of more diverse posterior beliefs $p_t^\star$. When observations $x_t \sim p(\cdot|z_t^\star)$ contain little information about the true state $z_t^\star$, posterior beliefs will be more uncertain, decreasing $\lambda_{\min}^{(a)}$, which falls to zero in the limit where posteriors $p_t^\star$ fail to span the space of possible beliefs (e.g. if some latent states are indistinguishable), making $B_{\mathrm{eq}}^{(a)}$ no longer full rank, and hence singular.[4] Lastly, the Z-dependence in Eq. (3.11) indicates that reward estimation is easier when the latent space is lower dimensional, in which case prior knowledge of the latent structure is more valuable.

3.4 Derivation of Theorem 1

We would like to bound the error in the action-wise, vector-valued mean reward estimators $\hat{\mu}^{(a)}$, defined in Eq. (3.8), and (as discussed in Section 3.2), used by Algorithm 2 with the linear bandit context vector c_t set equal to the vector of posterior probabilities over the latent state, $\hat{p}_t$. As stated in Theorem 1, we set $(\theta, \Phi) = (\theta^\star, \Phi^\star)$ throughout this section,

[4] Furthermore, if an action a is rarely or never optimal, $B_{\mathrm{eq}}^{(a)}$ will approach the zero matrix, and again $\lambda_{\min}^{(a)} \to 0$ and the bound becomes weak due to less data for action a.

and thus replace the model posterior $\hat{p}_t$ with the "true" posterior $p^\star_t$ as defined in Eq. (3.4). We will occasionally denote the T-dependence of some quantities explicitly as an argument, when it is helpful to remember, but will in general leave it suppressed in the interest of simplicity.

It will be useful to express the difference between the estimated (Eq. (3.8)) and true mean reward parameters as

$$\hat{\mu}^{(a)} - \mu^{(a)}_\star = (B^{(a)})^{-1} g^{(a)}, \tag{3.14}$$

where

$$g^{(a)} := f^{(a)}_\mu - B^{(a)}\mu^{(a)}_\star = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p^\star_t \left(r_t - (p^\star_t)^\top \mu^{(a)}_\star\right). \tag{3.15}$$

For reference, it is also useful to write down the element-wise definitions of the vector $g^{(a)}$ and matrix $B^{(a)}$:

$$g^{(a)}_z = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p^\star(z_t = z|x_{1:t}) \Big(r_t - \sum_{z'} p^\star(z_t = z'|x_{1:t})(\mu^{(a)}_\star)_{z'}\Big). \tag{3.16}$$

$$B^{(a)}_{zz'} = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p^\star(z_t = z|x_{1:t}) p^\star(z_t = z'|x_{1:t}). \tag{3.17}$$

We will drop the * superscript on p_t in the following sections, to avoid notational clutter, but emphasize that throughout this section, all quantities are conditioned on the true parameters $(\theta^\star, \Phi^\star)$. We will also occasionally use the shorthand notation

$$p_{t:t'}(z) := p(z_{t'} = z|x_{t:t'}) \tag{3.18}$$

to simplify expressions. For simplicity, we will remove the * when denoting the transition matrix; we restore it in Theorem 1.

The derivation of Theorem 1 proceeds as follows. In Appendix 3.4.1 we derive several intermediate results using a contraction property [15] of the Kullback-Leibler divergence between two posterior beliefs over the state of a hidden Markov process, which implies that the KL distance between two beliefs about the current latent state z_t contracts exponentially in time as the beliefs are updated over time with additional context observations x_t. We use this result to upper bound the dependence of posterior beliefs of the form $p_t(z) := p(z_t = z|x_{1:t})$ on data $x_{t-\tau}$ observed in the distant past (large τ), such that probabilities $p_t(z)$ and $p_{t'}(z)$ may be treated as approximately i.i.d. random variables when $|t' - t|$ is large. Since the estimators $\hat{\mu}^{(a)}$ are constructed via linear regression with probabilities $p_t(z)$ being dependent variables, the approximate i.i.d. nature of time-separated

posteriors leads to a reduction (and asymptotic convergence to zero) in estimator variance. We demonstrate this explicitly as follows: In Appendix 3.4.2 and Appendix 3.4.3, we use the results of Appendix 3.4.1 to obtain element-wise upper bounds on the variance of, respectively, the error vector $g^{(a)}$ and the empirical inverse covariance matrix $B^{(a)}$. In Appendix 3.4.4 we convert the element-wise bound on $B^{(a)}$ into a bound on the largest eigenvalue of $(B^{(a)})^{-1}$, which we use in Appendix 3.4.5 along with the bound on $g^{(a)}$ to obtain the final high-probability bound on the estimator error $\hat{\mu}^{(a)} - \mu_\star^{(a)}$.

3.4.1 Mixing rate bounds on conditional posterior probabilities

In this section we will derive an upper bound on the expected total variation distance, $\mathbb{E}[\sum_z |p_t(z) - q_t(z)|]$ and KL divergence $D_{KL}[p_t(z)||q_t(z)]$, between two distinct posteriors (p_t, q_t) obtained by updating corresponding priors (p_1, q_1) with the same sequence of context observations $x_{1:t}$, and using the same likelihood function and transition matrix. The contraction of these distribution distances indicates that the posterior probabilities at a given time depend predominantly on recent observations, with dependence on distant past observations, $x_{t-\tau}$, being exponentially suppressed (with respect to τ).

As stated in Theorem 1, we assume that the latent Markov process is ergodic, and thus has a unique equilibrium distribution (or stationary distribution) $\rho_{\text{eq}}^{(\phi)}(z)$ defined by $\Phi\rho_{\text{eq}}^{(\phi)} = \rho_{\text{eq}}^{(\phi)}$.

Our analysis will make use of the *minimal mixing rate* [15] of a transition matrix,

$$\gamma_\phi := \min_{z_1, z_2} \sum_z \min(\phi_{z,z_1}, \phi_{z,z_2}). \tag{3.19}$$

Given two initial distributions $p_1(z) = \mathbf{1}(z = z_1)$ and $p_2(z) = \mathbf{1}(z = z_2)$, with all of their probability mass concentrated respectively on states z_1 and z_2, the quantity $\sum_z \min(\phi_{z,z_1}, \phi_{z,z_2})$ is the minimal probability mass which is moved to shared successor states z by applying the transition matrix to p_1 and p_2. Thus, γ_ϕ quantifies the minimal probability mass that is moved from different states to a shared state, for any initial distributions p_1 and p_2. The minimal mixing rate was used by [15] to prove a contraction theorem for the KL divergence between two different distributions:

Theorem 1 (Theorem 3 in [15])**.** *For any two prior distributions p_0 and q_0 over states $z \in \{1, ..., Z\}$, the distributions $p = \Phi p_0$, $q = \Phi q_0$ induced by a transition matrix Φ satisfy*

$$D_{KL}[p||q] \leq (1 - \gamma_\phi) D_{KL}[p_0||q_0], \tag{3.20}$$

with the minimal mixing rate γ_ϕ defined in Eq. (3.19).

We will also make use of the fact [15] that conditioning on additional data reduces the KL divergence between different distributions, in expectation:

Lemma 1.1. *Given two distinct priors $p(z)$ and $q(z)$, and corresponding posteriors obtained by conditioning on a real-valued observation x generated from a known likelihood distribution $\ell(x|z)$,*

$$p(z|x) = p(z)\ell(x|z)/p(x), \quad q(z|x) = q(z)\ell(x|z)/q(x), \tag{3.21}$$

where $p(x) := \sum_z p(z)\ell(x|z)$ and $q(x) := \sum_z q(z)\ell(x|z)$, the KL divergence between the posteriors $p(z|x)$ and $q(z|x)$ satisfies

$$\mathbb{E}_{x\sim p(x)}[D_{KL}[p(z|x)||q(z|x)]] \leq D_{KL}[p(z)||q(z)]. \tag{3.22}$$

Proof. Using Eq. (3.21), we have

$$\begin{aligned}
\mathbb{E}_{x\sim p(x)}[D_{KL}[p(z|x)||q(z|x)]] &= \mathbb{E}_{x\sim p(x)}\left[\sum_z \frac{p(z)\ell(x|z)}{p(x)} \log\left(\frac{p(z)\ell(x|z)}{p(x)}\frac{q(x)}{q(z)\ell(x|z)}\right)\right] \\
&= \mathbb{E}_{x\sim p(x)}\left[\sum_z \frac{p(z)\ell(x|z)}{p(x)}\left(\log\frac{p(z)}{q(z)} - \log\frac{p(x)}{q(x)}\right)\right] \\
&= \sum_z p(z)\log\frac{p(z)}{q(z)}\,\mathbb{E}_{x\sim p(x)}[\ell(x|z)/p(x)] - \mathbb{E}_{x\sim p(x)}\left[\frac{\sum_z p(z)\ell(x|z)}{p(x)}\log\frac{p(x)}{q(x)}\right] \\
&= D_{KL}[p(z)||q(z)] - D_{KL}[p(x)||q(x)].
\end{aligned} \tag{3.23}$$

In the last line, we have used the fact that $\frac{\sum_z p(z)\ell(x|z)}{p(x)} = 1$ by definition, and $\mathbb{E}_{x\sim p(x)}[\ell(x|z)/p(x)] = \mathbb{E}_{x\sim\ell(\cdot|z)}[1] = 1$. Since $D_{KL}[p(x)||q(x)] \geq 0$, we recover Eq. (3.22). □

Eq. (3.20) and Eq. (3.22) can be combined to show that the KL divergence between two prior beliefs over the hidden state contracts in expectation during a single transition and subsequent observation:

Lemma 1.2. *Given two prior probability distributions $q_0(z)$ and $\tilde{q}_0(z)$ over the hidden state z, the posterior distributions over the successor state z', conditional on observing $x \sim p(\cdot|z';\theta)$, that is*

$$q(z') \propto \sum_z \Phi_{z',z} q_0(z) p(x|z';\theta), \quad \tilde{q}(z') \propto \sum_z \Phi_{z',z} \tilde{q}_0(z) p(x|z';\theta),$$

where the sequence $x_{1:t}$ is generated via a sequence of latent states using the transition matrix Φ, satisfy

$$\mathbb{E}_{x\sim p(\cdot|z';\theta),z'\sim\Phi\tilde{q}_0}[D_{KL}[\tilde{q}||q]] \leq (1-\gamma_\phi)D_{KL}[\tilde{q}_0||q_0], \tag{3.24}$$

where the expectation is taken over $x \sim p(x) = \sum_{z,z'} \Phi_{z',z}\tilde{q}_0(z)p(x|z';\theta)$.

Proof. Applying Eq. (3.22) with prior probability vectors $\Phi\tilde{q}$ and Φq over z_t, we have

$$\mathbb{E}_{x\sim p(x)}[D_{KL}[\tilde{q}||q]] \leq D_{KL}[\Phi\tilde{q}_0||\Phi q_0].$$

where $p(x) = \sum_{z'}(\Phi\tilde{q}_0)_{z'}p(x|z';\theta)$. Applying Eq. (3.20), we recover Eq. (3.24). □

Note that Eq. (3.24) – and consequently also Eqs. (3.27) and (3.30) below – is asymmetric with respect to q and $\tilde{q}$, since the expectation is over data x generated with the first argument, $\tilde{q}_0$.

Eq. (3.24) can be applied recursively to show that the KL divergence contracts exponentially as the two distributions are propagated forward in time:

Lemma 1.3. *Given two prior probability distributions $q_0(z)$ and $\tilde{q}_0(z)$ over the initial latent state z_0, the resulting posterior distributions over the state z_t at time t, that is*

$$q_t(z') := \sum_z q_0(z)p(z_t = z'|x_{1:t}, z_0 = z), \tag{3.25}$$

$$\tilde{q}_t(z') := \sum_z \tilde{q}_0(z)p(z_t = z'|x_{1:t}, z_0 = z), \tag{3.26}$$

satisfy

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}[D_{KL}[\tilde{q}_t||q_t]] \leq e^{-\gamma_\phi t}D_{KL}[\tilde{q}_0||q_0], \tag{3.27}$$

where the expectation is over histories $x_{1:t}$ which are generated from initial latent states $z_0 \sim \tilde{q}_0(\cdot)$.

Proof. Applying Eq. (3.24) to the transition at time t, with priors $(\tilde{q}_0, q_0) \to (\tilde{q}_{t-1}, q_{t-1})$ in Eq. (3.24) determined by a fixed sequence $x_{1:t-1}$ of preceding data, we have

$$\mathbb{E}_{x_t|x_{1:t-1},z_0\sim\tilde{q}_0}[D_{KL}[\tilde{q}_t||q_t]] \leq (1-\gamma_\phi)D_{KL}[\tilde{q}_{t-1}||q_{t-1}], \tag{3.28}$$

where we have denoted that the expectation is taken only over $x_t \sim p(x) = \sum_z(\Phi\tilde{q}_{t-1})_z p(x|z;\theta)$. Taking the remaining expectations recursively over $x_{t-1}, ..., x_1$, backwards in time, we have

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}[D_{KL}[\tilde{q}_t||q_t]] \leq (1-\gamma_\phi)^t D_{KL}[\tilde{q}_0||q_0], \tag{3.29}$$

Since $(1-\gamma_\phi)^t = (e^{\log(1-\gamma_\phi)})^t = e^{t\log(1-\gamma_\phi)} < e^{-\gamma_\phi t}$ for $\gamma_\phi \in (0,1)$ and $t > 0$, we recover Eq. (3.27). □

Note that Eq. (3.27) is a conservative bound, for two reasons: (1) If there exist pairs of states (z_1, z_2) in Eq. (3.19) – e.g. spatially distant states – which cannot transition to any common state z, we have $\gamma_\phi = 0$. However, mixing may still occur efficiently over several timesteps – e.g. allowing for several transitions between spatially connected states – leading to a similar exponential contraction with respect to a more general mixing rate. (2) Eq. (3.22) is a weaker bound than Eq. (3.23), which may be substantially tighter when the marginal context distributions $p(x)$ and $q(x)$ are separated by a large KL distance. This can occur when the conditional context distributions $p(x|z;\theta)$ – denoted $\ell(x|z)$ in Lemma 1.1 – are very different, making observations x highly informative about z.

Eq. (3.27) can be converted into a bound on the expected total variation distance, or 1-norm, between two posteriors:

Corollary 1.1. *The 1-norm difference between two distributions $(\tilde{q}_t, q_t)$ over the state z_t, as defined in Eqs. (3.25)-(3.26), satisfies the upper bound*

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}\left[\sum_z |\tilde{q}_t(z) - q_t(z)|\right] \leq e^{-\frac{1}{2}\gamma_\phi t}\sqrt{2D_{KL}[\tilde{q}_0||q_0]}. \tag{3.30}$$

Proof. Pinsker's inequality states that for any two probability distributions $\tilde{q}$ and q, the 1-norm and KL divergence satisfy $||\tilde{q}-q||_1 \leq \sqrt{2D_{KL}[\tilde{q}||q]}$.[5] Setting $||\tilde{q}-q||_1 = \sum_z |\tilde{q}_t(z)-q_t(z)|$, and taking the expectation, we have

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}\left[\sum_z |\tilde{q}_t(z) - q_t(z)|\right] \leq \mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}\left[\sqrt{2D_{KL}[\tilde{q}_t||q_t]}\right].$$

Applying Jensen's inequality to bring the expectation under the square root, we have

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}\left[\sum_z |\tilde{q}_t(z) - q_t(z)|\right] \leq \sqrt{2\cdot\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}[D_{KL}[\tilde{q}_t||q_t]]}.$$

Applying Eq. (3.27), we arrive at Eq. (3.30). □

3.4.2 Partial bound on the estimator error

In this section, we compute the variance of the vector $g^{(a)}$ defined in Eq. (3.15), across different reward and observation histories. We show that the variance converges to zero

[5]The symmetry of the left hand side under exchange of $\tilde{q}$ and q implies the same relation holds with respect to the reverse KL divergence $D_{KL}[q||\tilde{q}]$.

as $T \to \infty$, and then show that $|g^{(a)}|$ converges to zero asymptotically. In the following sections, we will use this result to bound the estimator error $\hat{\mu}^{(a)} - \mu_\star^{(a)} = (B^{(a)})^{-1} g^{(a)}$.

Lemma 1.4. *When the ground truth parameters (θ, Φ) are known, each element of $g^{(a)}$, Eq. (3.15), satisfies the upper bound*

$$(g_z^{(a)}/T)^2 \leq \frac{1}{\delta \cdot T}\Big(\sigma_{\text{eq}}^2 + ||\mu_\star^{(a)}||_1^2 \frac{4}{\gamma_\phi}\big(1 + \log c_\phi\big)\Big) \tag{3.31}$$

with probability at least $1 - \delta$, for any $\delta \in (0, 1)$, where

$$\sigma_{\text{eq}}^2 := \max_a \sum_z \rho_{\text{eq}}^{(\phi)}(z) \text{Var}[r|z, a] \tag{3.32}$$

is the maximal variance in rewards when the latent state has reached equilibrium, and $\gamma_\phi^{-1} \log c_\phi := \tau^\star$ is the integer number of timesteps satisfying

$$\tau^\star := \min_{\tau \in \mathbb{N}} |\log D_\phi(\tau) - \gamma_\phi \tau|, \tag{3.33}$$

where

$$D_\phi(\tau) := \max_z \max_{t \geq 1} \mathbb{E}_{x_{1:t+\tau}}\left[D_{KL}[p(z_{t+\tau}|x_{1:t+\tau})||p(z_{t+\tau}|z_t = z; x_{1:t+\tau})]\right] \tag{3.34}$$

is a measure of how much information the latent state z_t at any time t can possibly contain about a future latent state $z_{t+\tau}$.

Proof. We will use the shorthand notation $\delta_t^{(a)} := \mathbf{1}(a_t = a)$ for the indicator function which picks out times t for a given action a. First, we observe that the expectation of $g^{(a)}$

(conditional on any action sequence $a_{1:T}$) is zero:

$$\begin{aligned}
\mathbb{E}[g_z^{(a)}|a_{1:T}] &= \mathbb{E}_{x_{1:T}}[g_z^{(a)}|x_{1:T}, a_{1:T}] \\
&= \mathbb{E}_{x_{1:T}}\Big[\sum_{t=1}^{T} \delta_t^{(a)} p(z_t = z|x_{1:t})\, \mathbb{E}[r_t|x_{1:T}, a_t = a] \\
&\qquad - \sum_{t=1}^{T} \delta_t^{(a)} p(z_t = z|x_{1:t}) \sum_{z'} p(z_t = z'|x_{1:t})(\mu_\star^{(a)}) r_{z'}\Big] \\
&= \sum_{t=1}^{T} \delta_t^{(a)}\, \mathbb{E}_{x_{1:T}}\left[p(z_t = z|x_{1:t}) \sum_{z'} (p(z_t = z'|x_{1:T}) - p(z_t = z'|x_{1:t}))(\mu_\star^{(a)})_{z'}\right] \\
&= \sum_{t=1}^{T} \delta_t^{(a)} \sum_{z'} (\mu_\star^{(a)})_{z'} \cdot \mathbb{E}_{x_{1:t}}\left[p(z_t = z|x_{1:t})(\mathbb{E}_{x_{t+1:T}}[p(z_t = z'|x_{1:T})] - p(z_t = z'|x_{1:t}))\right] \\
&= \sum_{t=1}^{T} \delta_t^{(a)} \sum_{z'} (\mu_\star^{(a)})_{z'} \cdot \mathbb{E}_{x_{1:t}}[p(z_t = z|x_{1:t})(p(z_t = z|x_{1:t}) - p(z_t = z|x_{1:t}))] = 0.
\end{aligned} \tag{3.35}$$

Here, we have used the fact that $\mathbb{E}[r_t|x_{1:T}, a_t = a] = \sum_{z'} p(z_t = z'|x_{1:T})(\mu_\star^{(a)})_{z'}$ to take the expectation over reward data, followed by the partial expectation over context data $x_{t+1:T}$.

Since $\mathbb{E}[g_z^{(a)}] = 0$, we compute the variance to obtain an upper bound on $|g_z^{(a)}|$. To compute the variance of the vector element $g_z^{(a)}$, we first take the expectation over rewards, conditional on a specific context history $x_{1:T}$. Defining the reward noise

$$\eta_t^{(a)} := r_t - \sum_{z'} p(z_t = z'|x_{1:t})(\mu_\star^{(a)})_{z'} = r_t - p_t^\top \mu_\star^{(a)}, \tag{3.36}$$

so that for brevity we can write $g^{(a)} = \sum_t \delta_t^{(a)} p_t \eta_t^{(a)}$, or equivalently

$$g_z^{(a)} = \sum_{t=1}^{T} \delta_t^{(a)} p(z_t = z|x_{1:t}) \eta_t^{(a)},$$

we have (for any (z_1, z_2))

$$\mathbb{E}[g_{z_1}^{(a)} g_{z_2}^{(a)}|x_{1:T}, a_{1:T}] = \sum_{t,t'} \delta_t^{(a)} \delta_{t'}^{(a)} p(z_t = z_1|x_{1:t}) p(z_{t'} = z_2|x_{1:t'}) \cdot \mathbb{E}[\eta_t^{(a)} \eta_{t'}^{(a)}|x_{1:T}, a_t = a_{t'} = a]. \tag{3.37}$$

Since $\mathbb{E}[r_t|x_{1:T}, a_t = a] = \sum_z p(z_t = z|x_{1:T})(\mu_\star^{(a)})_z$ and $\mathbb{E}[r_t r_{t'}|x_{1:T}, a_t = a_{t'} = a] = \sum_{z,z'} p(z_t = z, z_{t'} = z'|x_{1:T})(\mu_\star^{(a)})_z(\mu_\star^{(a)})_{z'}$, the correlation between reward noise at times t and $t' \neq t$ is

$$\begin{aligned}
&(\text{for } t \neq t') \\
&\mathbb{E}[\eta_t \eta_{t'}|x_{1:T}, a_t = a_{t'} = a] = \\
&\quad \sum_{z,z'} (\mu_\star^{(a)})_z (\mu_\star^{(a)})_{z'} \big[p(z_t = z, z_{t'} = z'|x_{1:T}) - p(z_t = z|x_{1:t}) p(z_{t'} = z'|x_{1:T}) \\
&\qquad - p(z_t = z|x_{1:T}) p(z_{t'} = z'|x_{1:t'}) + p(z_t = z|x_{1:t}) p(z_{t'} = z'|x_{1:t'}) \big].
\end{aligned} \tag{3.38}$$

When $t = t'$ we have

$$\begin{aligned}
\mathbb{E}[\eta_t^2|x_{1:T}, a_t = a] = &\sum_z p(z_t = z|x_{1:T})((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) \\
&- 2\Big(\sum_z p(z_t = z|x_{1:t})(\mu_\star^{(a)})_z \Big) \Big(\sum_{z'} p(z_t = z'|x_{1:T})(\mu_\star^{(a)})_{z'} \Big) \\
&+ \Big(\sum_z p(z_t = z|x_{1:t})(\mu_\star^{(a)})_z \Big)^2,
\end{aligned} \tag{3.39}$$

where

$$\sigma_z^{(a)} := \mathbb{E}_{r \sim p(\cdot|z,a)}[r^2] - \mathbb{E}_{r \sim p(\cdot|z,a)}[r]^2 = \mathbb{E}_{r \sim p(\cdot|z,a)}[r^2] - [(\mu_\star^{(a)})_z]^2. \tag{3.40}$$

We now take the expectation over $x_{1:T}$. Because Eq. (3.37) only depends on $x_{t'+1:T}$ via the conditional expectation of reward noise $\mathbb{E}[\eta_t \eta_{t'}|x_{1:T}]$, we can take the partial expectation over $x_{t'+1:T}$ as follows:

$$\begin{aligned}
&\mathbb{E}[g_{z_1}^{(a)} g_{z_2}^{(a)}|a_{1:T}] = \mathbb{E}_{x_{1:T}} \left[\mathbb{E}[g_{z_1}^{(a)} g_{z_2}^{(a)}|x_{1:T}, a_{1:T}] \right] \\
&\quad = 2 \sum_{t,t'>t} \delta_t^{(a)} \delta_{t'}^{(a)} \, \mathbb{E}_{x_{1:t'}} \left[p(z_t = z_1|x_{1:t}) p(z_{t'} = z_2|x_{1:t'}) \cdot \mathbb{E}_{x_{t'+1:T}}[\mathbb{E}[\eta_t \eta_{t'}|x_{1:T}, a_t = a_{t'} = a]] \right] \\
&\qquad + \sum_t \delta_t^{(a)} \, \mathbb{E}_{x_{1:t}} \left[p(z_t = z_1|x_{1:t}) p(z_t = z_2|x_{1:t}) \cdot \mathbb{E}_{x_{t+1:T}}[\mathbb{E}[\eta_t^2|x_{1:T}, a_t = a]] \right]
\end{aligned} \tag{3.41}$$

where we have decomposed the double sum over time as $\sum_{t,t'} = \sum_{t=t'} + 2\sum_{t,t'>t}$. Using

Eq. (3.38) for the $t < t'$ terms, we have

$$\begin{aligned}
(\text{for } t < t') \quad & \mathbb{E}_{x_{t'+1:T}}[\mathbb{E}[\eta_t \eta_{t'} | x_{1:T}, a_t = a_{t'} = a]] \\
&= \sum_{z,z'} (\mu_\star^{(a)})_z (\mu_\star^{(a)})_{z'} \big[\mathbb{E}_{x_{t'+1:T}}[p(z_t = z, z_{t'} = z' | x_{1:T})] \\
&\qquad - p(z_t = z | x_{1:t}) \, \mathbb{E}_{x_{t'+1:T}}[p(z_{t'} = z' | x_{1:T})] \\
&\qquad - \mathbb{E}_{x_{t'+1:T}}[p(z_t = z | x_{1:T})] p(z_{t'} = z' | x_{1:t'}) \\
&\qquad + p(z_t = z | x_{1:t}) p(z_{t'} = z' | x_{1:t'}) \big] \\
&= \sum_{z,z'} (\mu_\star^{(a)})_z (\mu_\star^{(a)})_{z'} \big[p(z_t = z, z_{t'} = z' | x_{1:t'}) - p(z_t = z | x_{1:t}) p(z_{t'} = z' | x_{1:t'}) \\
&\qquad - p(z_t = z | x_{1:t'}) p(z_{t'} = z' | x_{1:t'}) + p(z_t = z | x_{1:t}) p(z_{t'} = z' | x_{1:t'}) \big] \\
&= \sum_{z,z'} (\mu_\star^{(a)})_z (\mu_\star^{(a)})_{z'} p(z_t = z | x_{1:t'}) \big(p(z_{t'} = z' | z_t = z, x_{1:t'}) - p(z_{t'} = z' | x_{1:t'}) \big)
\end{aligned} \tag{3.42}$$

In the second equality we have cancelled two equivalent terms, and in the last line we have factored the joint distribution over (z, z') into a marginal and conditional. Similarly, using Eq. (3.39) for the $t' = t$ terms, we have

$$\begin{aligned}
\mathbb{E}[\eta_t^2 | x_{1:t}, a_t = a] &= \mathbb{E}_{x_{t+1:T}}[\mathbb{E}[\eta_t^2 | x_{1:T}, a_t = a]] \\
&= \sum_z p(z_t = z | x_{1:t}) ((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) - \Big(\sum_z p(z_t = z | x_{1:t}) (\mu_\star^{(a)})_z \Big)^2 \\
&\le \sum_z p(z_t = z | x_{1:t}) ((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2).
\end{aligned} \tag{3.43}$$

Substituting Eqs. (3.42) and (3.43) into Eq. (3.41), taking the absolute value to obtain an upper bound, using $p(z_t = z_1 | x_{1:t}) p(z_{t'} = z_2 | x_{1:t'}) \le 1$ and $p(z_t = z | x_{1:t'}) \le 1$ to simplify the expression, using the fact that $\mathbb{E}_{x_{1:t}}[p(z_t = z | x_{1:t})] = \rho_{\text{eq}}^{(\phi)}(z)$ in the $t = t'$ contribution, and setting $z_1 = z_2$ for simplicity, we have

$$\begin{aligned}
\text{Var}[g_{z_1}^{(a)}] \le T \sum_z \rho_{\text{eq}}^{(\phi)}(z) ((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) \\
+ \sum_{z,z'} |(\mu_\star^{(a)})_z (\mu_\star^{(a)})_{z'}| \times 2 \sum_{t,t'>t} \mathbb{E}_{x_{1:t'}} \left[|p(z_{t'} = z' | z_t = z, x_{1:t'}) - p(z_{t'} = z' | x_{1:t'})| \right]
\end{aligned} \tag{3.44}$$

We have also used $\delta_t^{(a)}\delta_{t'}^{(a)} \leq 1$ and have removed the action-conditioning on $\mathrm{Var}[g^{(a)}]$, since after setting $\delta_t^{(a)}\delta_{t'}^{(a)} \leq 1$ the right-hand side no longer depends on the action sequence, and thus the inequality holds for any action sequence. Introducing a free parameter τ_1 satisfying $1 \leq \tau_1 \leq t' - t$, we take the partial expectation over $x_{t+\tau_1:t'}$ of the difference in conditional probabilities by applying Corollary 1.1 to bound the expectation value over $x_{t+\tau_1+1:t'}$:

$$\begin{aligned}
&\mathbb{E}_{x_{1:t'}}[|p(z_{t'} = z'|z_t = z, x_{1:t'}) - p(z_{t'} = z'|x_{1:t'})|] \\
&\quad = \mathbb{E}_{x_{1:t+\tau_1}}\, \mathbb{E}_{x_{t+\tau_1+1:t'}}[|p(z_{t'} = z'|z_t = z, x_{1:t'}) - p(z_{t'} = z'|x_{1:t'})|] \\
&\quad \leq e^{-\frac{1}{2}\gamma_\phi(t'-(t+\tau_1))}\, \mathbb{E}_{x_{1:t+\tau_1}}\left[\sqrt{2D_{KL}[p(z_{t+\tau_1}|x_{1:t+\tau_1})||p(z_{t+\tau_1}|z_t = z; x_{1:t+\tau_1})]}\right] \\
&\quad \leq e^{-\frac{1}{2}\gamma_\phi(t'-(t+\tau_1))}\sqrt{2 \cdot \mathbb{E}_{x_{1:t+\tau_1}}\left[D_{KL}[p(z_{t+\tau_1}|x_{1:t+\tau_1})||p(z_{t+\tau_1}|z_t = z; x_{1:t+\tau_1})]\right]} \\
&\quad \leq e^{-\frac{1}{2}\gamma_\phi(t'-(t+\tau_1))}\sqrt{2D_\phi(\tau_1)}.
\end{aligned}$$

In the second inequality, we have applied Jensen's inequality to bring the expectation inside the square root. In the last line, we have recalled the definition of $D_\phi(\tau_1)$ in Eq. (3.34). For $\tau_1 \gg 1/\gamma_\phi$, the latent state will have evolved through multiple mixing times, so we expect $D_\phi(\tau_1)$ to become small, decreasing to zero as $\tau_1 \to \infty$.

We now introduce a second free parameter $\tau_0 \in \mathbb{N}$ (which we will optimize below), and use it to decompose the sum over $t' - t$ into a contribution from widely separated times, $t' - t > \tau_0$, where the exponential suppression is strong, and a contribution from nearby times, $t' - t \leq \tau_0$, over which the posterior probabilities may be more strongly correlated and there is not significant exponential suppression:

$$\begin{aligned}
\mathrm{Var}[g_{z_1}^{(a)}] \leq T\sum_z \rho_{\mathrm{eq}}^{(\phi)}(z)((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) \\
+ 2||\mu_\star^{(a)}||_1^2 \sum_{t,t'>t}\left[\mathbf{1}(t'-t \leq \tau_0) + \mathbf{1}(t'-t > \tau_0)e^{-\frac{1}{2}\gamma_\phi(t'-(t+\tau_1))}\sqrt{2D_\phi(\tau_1)}\right].
\end{aligned} \tag{3.45}$$

Here, we have used the fact that

$$\sum_{z,z'}|(\mu_\star^{(a)})_z(\mu_\star^{(a)})_{z'}| \leq \sum_z |(\mu_\star^{(a)})_z| \times \sum_{z'}|(\mu_\star^{(a)})_{z'}| = ||\mu_\star^{(a)}||_1^2,$$

and (in the $t' - t \leq \tau_0$ term) the fact that the difference of probabilities in Eq. (3.44) is between 0 and 1. The $t' - t > \tau_0$ contribution can be upper bounded as follows:

$$\sum_{t,t'>t}\mathbf{1}(t'-t > \tau_0)e^{-\frac{1}{2}\gamma_\phi(t'-(t+\tau_1))} \leq T\sum_{\tau=\tau_0+1}^{T} e^{-\frac{1}{2}\gamma_\phi(\tau-\tau_1)} \leq T\int_{\tau_0}^{\infty} d\tau e^{-\frac{1}{2}\gamma_\phi(\tau-\tau_1)} = \frac{2T}{\gamma_\phi}e^{-\frac{1}{2}\gamma_\phi(\tau_0-\tau_1)}.$$

Here, we have used monotonicity with respect to τ to bound the discrete sum with a continuous integral. Using this in Eq. (3.45), we have

$$\mathrm{Var}[g_{z_1}^{(a)}] \leq T \sum_z \rho_{\mathrm{eq}}^{(\phi)}(z)((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) + 2||\mu_\star^{(a)}||_1^2 \Big(T\tau_0 + \frac{2T}{\gamma_\phi} e^{-\frac{1}{2}\gamma_\phi(\tau_0-\tau_1)} \sqrt{2D_\phi(\tau_1)} \Big). \tag{3.46}$$

Setting to zero the derivative with respect to τ_0, and solving for τ_0, we find the optimal value

$$\tau_0^\star := \tau_1 + \frac{1}{\gamma_\phi} \log(2D_\phi(\tau_1)),$$

for which the upper bound becomes

$$\mathrm{Var}[g_{z_1}^{(a)}] \leq T \sum_z \rho_{\mathrm{eq}}^{(\phi)}(z)((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) + 2T||\mu_\star^{(a)}||_1^2 \Big(\tau_1 + \frac{1}{\gamma_\phi}\big(2 + \log D_\phi(\tau_1)\big) \Big).$$

We now approximately optimize τ_1 by setting it equal to the value $\tau_1^\star$ at which $\gamma_\phi \tau_1^\star = \log D_\phi(\tau_1^\star) := \log c_\phi$. Furthermore, since $\sum_z \rho_{\mathrm{eq}}^{(\phi)}(z)((\mu_\star^{(a)})_z)^2 < \sum_z ((\mu_\star^{(a)})_z)^2 = ||\mu_\star^{(a)}||_2^2 < ||\mu_\star^{(a)}||_1^2$ and $1/\gamma_\phi \geq 1$, the expression for $\mathrm{Var}[g_z^{(a)}]$ simplifies to:

$$\mathrm{Var}[g_z^{(a)}] \leq T\Big(\sigma_{\mathrm{eq}}^2 + ||\mu_\star^{(a)}||_1^2 \frac{4}{\gamma_\phi}\big(1 + \log c_\phi\big) \Big), \tag{3.47}$$

where

$$\sigma_{\mathrm{eq}}^2 := \max_a \sum_z \rho_{\mathrm{eq}}^{(\phi)}(z)(\sigma_z^{(a)})^2. \tag{3.48}$$

Finally, we apply Chebyshev's inequality, which states that

$$|g_z^{(a)} - \mathbb{E}[g_z^{(a)}]| < \sqrt{\frac{\mathrm{Var}[g_z^{(a)}]}{\delta}} \tag{3.49}$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$. Recalling from Eq. (3.35) that $\mathbb{E}[g_z^{(a)}] = 0$, we recover Eq. (3.31) above. □

3.4.3 Bound on the inverse covariance matrix

In this section we derive a theoretical bound on the action-wise inverse covariance matrix $B^{(a)}$ in the $T \to \infty$ limit.

We will (i) use a mild assumption on the frequency with which optimal actions are selected in order to lower bound the expected elements $\mathbb{E}[B^{(a)}_{z,z'}]$ of the action-wise inverse covariance matrices, (ii) show that the variance around this expectation decreases as $1/\gamma_\phi T$, and (iii) combine these results to obtain a high-probability lower bound on the empirical inverse covariance matrix $B^{(a)}$.

Recalling that the context history $x_{1:t}$ determines (conditional on the true task parameters[6]) an optimal action

$$a^\star_t := \arg\max_a \sum_z p^\star(z_t = z|x_{1:t})\mu^{(a)}_\star, \tag{3.50}$$

we state the lower bound of point (i) above:

Lemma 1.5. *Assuming that at any t the optimal action given $x_{1:t}$, Eq. (3.50), is selected by a policy π with probability at least $\pi_{\min} > 0$, the expectation over histories $x_{1:T}$ of the empirical inverse covariance matrix, $B^{(a)}$, satisfies the lower bound*

$$\frac{1}{T}\,\mathbb{E}[B^{(a)}(T)] \succcurlyeq \pi_{\min}\Sigma^{(a)}(T), \tag{3.51}$$

where $A \succcurlyeq B$ indicates that $A - B$ is positive semidefinite, and

$$\Sigma^{(a)}_{zz'}(T) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_{1:t}}\left[\mathbf{1}(a = a^\star_t)p(z_t = z|x_{1:t})p(z_t = z'|x_{1:t})\right]. \tag{3.52}$$

Proof. We first express the expectation value of the matrix element $B^{(a)}_{zz'}$ as a sum over expected values at each time,

$$\begin{aligned}\mathbb{E}[B^{(a)}_{zz'}(T)] &= \sum_{t=1}^{T} \mathbb{E}_{x_{1:t}}\left[\mathbb{E}_{r_{1:t-1},a_{1:t-1}|x_{1:t}}[\mathbf{1}(a_t = a)]p(z_t = z|x_{1:t})p(z_t = z'|x_{1:t})\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{x_{1:t}}\left[P_\pi(a_t = a|x_{1:t})p(z_t = z|x_{1:t})p(z_t = z'|x_{1:t})\right]. \end{aligned} \tag{3.53}$$

In the first line, we have decomposed the expectation into an inner context-conditioned expectation over actions and rewards, and an outer expectation over contexts. The former only involves the binary indicator $\mathbf{1}(a_t = a)$, and is the probability

$$P_\pi(a_t = a|x_{1:t}) := \mathbb{E}_{r_{1:t-1},a_{1:t-1}|x_{1:t}}[\mathbf{1}(a_t = a)] \tag{3.54}$$

[6] We restore the * notation in Eq. (3.50) to denote this.

that a given policy π selects action $a_t = a$ conditional on the context history $x_{1:t}$. As stated in Theorem 1, we make the mild assumption that the optimal action $a_t^\star$ is selected with a minimal nonzero probability $\pi_{\min}$. (Any policy that learns the task should converge to $\pi_{\min} \to 1$ as $T \to \infty$.) That is,

$$P_\pi(a_t = a|x_{1:t}) \geq \pi_{\min} \cdot \mathbf{1}(a = a_t^\star), \tag{3.55}$$

where we conservatively lower bound the probability at zero for $a \neq a_t^\star$. Since the rank one matrix $p_t p_t^\top$ with elements

$$(p_t p_t^\top)_{z,z'} = p(z_t = z|x_{1:t})p(z_t = z'|x_{1:t})$$

is positive semidefinite[7] for any $x_{1:t}$, Eq. (3.55) implies that, for any p_t,

$$P_\pi(a_t = a|x_{1:t})p_t p_t^\top \succcurlyeq \pi_{\min} \cdot \mathbf{1}(a = a_t^\star)p_t p_t^\top$$

and hence

$$\mathbb{E}_{x_{1:t}}[P_\pi(a_t = a|x_{1:t})p_t p_t^\top] \succcurlyeq \pi_{\min}\, \mathbb{E}_{x_{1:t}}[\mathbf{1}(a = a_t^\star)p_t p_t^\top].$$

Applying this bound to each matrix term of $\mathbb{E}[B_{zz'}^{(a)}(T)]$ in Eq. (3.53) we see that

$$\mathbb{E}[B^{(a)}(T)] \succcurlyeq \pi_{\min} \cdot T \cdot \Sigma^{(a)}(T), \tag{3.56}$$

with $\Sigma^{(a)}(T)$ defined in Eq. (3.52). Hence we recover the matrix lower bound Eq. (3.51) above. □

We now show that the variance of the empirical matrix B around its asymptotic expected form can be upper bounded:

Lemma 1.6. *When the ground truth parameters (θ, Φ) are known, the variance across histories $x_{1:T}$ of the empirical inverse covariance matrix element $B_{zz'}(T)$, satisfies the upper bound*

$$\mathrm{Var}\left[\frac{1}{T}B_{zz'}^{(a)}(T)\right] \leq \frac{2}{\gamma_\phi T}(c + \log\log(1/\rho_{\min})), \tag{3.57}$$

where $c \approx 6.78$, and $\rho_{\min} := \min_z \rho_{\mathrm{eq}}^{(\phi)}(z)$ is the equilibrium probability of the least probable latent state.

[7] This matrix has $Z - 1$ zero eigenvalues, and a nonzero eigenvalue $\sum_z p(z_t = z|x_{1:t})^2$.

Proof. The variance over context histories $x_{1:T}$ of the matrix element $B_{zz'}(T)$, conditioned on actions $a_{1:T}$ (and using the shorthand notation $\delta_t^{(a)} = \mathbf{1}(a_t = a)$), is

$$\begin{aligned}\mathrm{Var}[B^{(a)}_{zz'}|a_{1:T}] &= \mathbb{E}_{x_{1:T}}[(B^{(a)}_{zz'})^2|a_{1:T}] - \mathbb{E}_{x_{1:T}}[B^{(a)}_{zz'}|a_{1:T}]^2 \\ &= \sum_{t,t'} \delta_t^{(a)}\delta_{t'}^{(a)} \Big(\mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')p_{1:t'}(z)p_{1:t'}(z')] - \mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')]\,\mathbb{E}_{x_{1:t'}}[p_{1:t'}(z)p_{1:t'}(z')] \Big).\end{aligned} \tag{3.58}$$

Here, we have trivially taken the expectation over $x_{t'+1:T}$. Using again the shorthand notation $p_{t:t'}(z) := p(z_{t'} = z|x_{t:t'})$ (with $p_{t:t'} \in \mathbb{R}^Z$ denoting the vector of probabilities), and defining

$$\delta p_t(z;\tau) := p_{1:t}(z) - p_{t-\tau+1:t}(z), \tag{3.59}$$

we can write, for $t' > t$,

$$p_{1:t'}(z)p_{1:t'}(z') = (p_{t+1:t'}(z') + \delta p_{t'}(z;t'-t))(p_{t+1:t'}(z') + \delta p_{t'}(z';t'-t)), \tag{3.60}$$

Using Corollary 1.1 to bound the expectation over $x_{t+1:t'}$, and using the fact that

$$D_{KL}[p_{1:t}||\rho^{(\phi)}_{\mathrm{eq}}] = \sum_z p_{1:t}(z)\log\left(\frac{p_{1:t}(z)}{\rho^{(\phi)}_{\mathrm{eq}}(z)}\right) \leq \sum_z p_{1:t}(z)\log(1/\rho_{\min}) = \log(1/\rho_{\min}),$$

we have

$$\mathbb{E}_{x_{1:t'}}\Big[\sum_z |\delta p_{t'}(z;\tau)|\Big] \leq e^{-\frac{1}{2}\gamma_\phi \tau}\,\mathbb{E}_{x_{1:t}}[\sqrt{2D_{KL}[p_{1:t}||\rho^{(\phi)}_{\mathrm{eq}}]} \leq e^{-\frac{1}{2}\gamma_\phi\tau}\sqrt{2\log(1/\rho_{\min})} := u(\tau). \tag{3.61}$$

Thus, for $t' > t$,

$$\begin{aligned}|\,\mathbb{E}_{x_{1:t'}}[p_{1:t'}(z)p_{1:t'}(z')] - \mathbb{E}_{x_{t+1:t'}}[p_{t+1:t'}(z)p_{t+1:t'}(z')]| &\leq \mathbb{E}_{x_{1:t'}}[|\delta p_{t'}(z;t'-t)|] + \mathbb{E}_{x_{1:t'}}[|\delta p_{t'}(z';t'-t)|] \\ &\quad + \mathbb{E}_{x_{1:t'}}[|\delta p_{t'}(z;t'-t)|\cdot|\delta p_{t'}(z';t'-t)|] \\ &\leq 3u(t'-t),\end{aligned} \tag{3.62}$$

where we have used $p_{t+1:t'} \leq 1$ and $|\delta p_{t'}| \leq 1$ to conservatively bound the expectation. Applying the decomposition in Eq. (3.60) again for the first term in Eq. (3.58), we have

$$\mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')p_{1:t'}(z)p_{1:t'}(z')] \leq \mathbb{E}_{x_{1:t}}[p_{1:t}(z)p_{1:t}(z')]\cdot\mathbb{E}_{t+1:t'}[p_{t+1:t'}(z)p_{t+1:t'}(z')] + 3u(t'-t). \tag{3.63}$$

Here we have used the fact that $p_{1:t}(z)p_{1:t}(z') \leq 1$ to simplify the last term. Combining Eq. (3.62) and (3.63), we have (for $t' > t$)

$$|\,\mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')p_{1:t'}(z)p_{1:t'}(z')] - \mathbb{E}_{x_{1:t}}[p_{1:t}(z)p_{1:t}(z')]\,\mathbb{E}_{x_{1:t'}}[p_{1:t'}(z)p_{1:t'}(z')]| \leq 6u(t'-t). \tag{3.64}$$

As in Lemma 1.4, we now introduce a free parameter τ_0, and break the sum in Eq. (3.58) into a contributions from small $|t'-t|$ (where the difference in Eq. (3.58) may be large but cannot exceed one) and large $|t'-t|$ (where the upper bound on the difference in Eq. (3.58) is strong). The variance $\mathrm{Var}[B_{z,z'}]$, Eq. (3.58), can then be upper bounded:

$$\mathrm{Var}[B^{(a)}_{zz'}|a_{1:T}] \leq \sum_{t,t'} \delta^{(a)}_t \delta^{(a)}_{t'} \left[\mathbf{1}(|t'-t| \leq \tau_0) + \mathbf{1}(|t'-t| > \tau_0) 6u(|t'-t|)\right]$$

Using $\delta^{(a)}_t \delta^{(a)}_{t'} \leq 1$ to apply the inequality for any action sequence $a_{1:T}$, and thus removing the action conditioning, we have

$$\mathrm{Var}[B^{(a)}_{zz'}] \leq \sum_{t,t'} \mathbf{1}(|t'-t| \leq \tau_0) + 2\sum_{t,t'} \mathbf{1}(t'-t > \tau_0) 6u(t'-t). \tag{3.65}$$

Here, we have also used the symmetry of Eq. (3.58) under exchange of t and t' to sum only over $t' > t$. The bound on $\mathrm{Var}[B_{z,z'}]$ becomes

$$\begin{aligned}
\mathrm{Var}[B^{(a)}_{zz'}] &\leq T(2\tau_0+1) + 12T \sum_{\tau=\tau_0+1}^{T} e^{-\frac{1}{2}\gamma_\phi \tau}\sqrt{2\log(1/\rho_{\min})} \\
&\leq T(2\tau_0+1) + 12T\sqrt{2\log(1/\rho_{\min})} \int_{\tau_0}^{\infty} d\tau e^{-\frac{1}{2}\gamma_\phi\tau} \\
&= T(2\tau_0+1) + 12T\sqrt{2\log(1/\rho_{\min})}\frac{2}{\gamma_\phi} e^{-\frac{1}{2}\gamma_\phi\tau_0}
\end{aligned}$$

where we have again used the monotonicity with respect to τ to bound the discrete sum with a continuous integral. We are now in a position to optimize the free parameter τ_0 to make the bound as tight as possible. Setting to zero the derivative with respect to τ_0, and solving for τ_0, we find the optimal value

$$\tau_0^\star := \frac{1}{\gamma_\phi}\log\left(72\log(1/\rho_{\min})\right), \tag{3.66}$$

for which the upper bound becomes

$$\begin{aligned}
\mathrm{Var}[B^{(a)}_{zz'}] &\leq T + 2\frac{T}{\gamma_\phi}\left[2 + \log\left(72\log(1/\rho_{\min})\right)\right] \\
&\leq 2\frac{T}{\gamma_\phi}(c + \log\log(1/\rho_{\min})),
\end{aligned}$$

where we have used the fact that $\gamma_\phi \leq 1$, and $c = \frac{5}{2} + \log 72 = \frac{5}{2} + 3\log 2 + 2\log 3 \approx 6.78$. □

Note that the unusual log-log dependence in Eq. (3.57) originates in the exponential contraction in Eq. 3.27, which suppresses an initial KL-distance that is already logarithmic in probabilities.

Finally, we apply Chebyshev's inequality to bound the deviation of the $B^{(a)}_{zz'}$ from its asymptotic expected value:

Lemma 1.7. *When the ground truth parameters (θ, Φ) are known, any matrix element of the empirical inverse covariance matrix $B^{(a)}(T)$, for any particular history $(x_{1:T}, a_{1:T})$ of contexts and actions, satisfies the inequality*

$$\frac{1}{T}|B^{(a)}_{zz'}(T) - \mathbb{E}[B^{(a)}_{zz'}(T)]| \leq \sqrt{\frac{1}{\delta}\frac{2}{\gamma_\phi T}(c + \log\log(1/\rho_{\min}))} \tag{3.67}$$

where $c \approx 6.78$, with probability at least $1 - \delta$, for any $\delta \in (0, 1)$.

Proof. Chebyshev's inequality states that for any random variable X with variance $\mathrm{Var}[X]$, $|X - \mathbb{E}[X]| \leq \sqrt{\mathrm{Var}[X]/\delta}$ with probability at least $1 - \delta$. Setting $X = \frac{1}{T}B^{(a)}_{zz'}$ and using Eq. (3.57) to upper bound the variance, we recover Eq. (3.67) above. □

3.4.4 Bound on covariance matrix eigenvalues

In Appendix 3.4.3 we derived a high-probability upper bound on the deviation of the elements of the empirical inverse covariance matrix $B^{(a)}$ from their asymptotic expected values. We would like to convert this into a bound on the covariance matrix $(B^{(a)})^{-1}$, in order to bound the estimator error $(B^{(a)})^{-1}g^{(a)}$, Eq. (3.14). In this section, we show that an element-wise bound such as Eq. (3.67) can be converted to an eigenvalue bound which can be applied to the inverse matrix.

Lemma 1.8. *For symmetric matrices $\bar{M}$, $M = \bar{M} + \Delta M$, with $|\Delta M_{z,z'}| \leq U_\delta$ for any given (z, z') with probability at least $1 - \delta$, the minimal eigenvalue λ_1 of M satisfies the lower bound*

$$\lambda_1 \geq \bar{\lambda}_1 - ZU_\delta \tag{3.68}$$

with probability at least $1 - Z\delta$, where $\bar{\lambda}_1$ is the minimal eigenvalue of $\bar{M}$.

Proof. Let λ_1 and $\bar{\lambda}_1$ be, respectively, the minimal eigenvalues of M and $\bar{M}$. Since M and $\bar{M}$ are symmetric, ΔM is also symmetric. The Weyl inequality for symmetric, real-valued square matrices states that if $\bar{\lambda}_1$ and $\lambda_1^{(\Delta)}$ are the minimal eigenvalues of matrices $\bar{M}$ and ΔM, then the minimal eigenvalue λ_1 of the matrix sum $\bar{M} + \Delta M$ satisfies the lower bound

$$\lambda_1 \geq \bar{\lambda}_1 + \lambda_1^{(\Delta)}. \tag{3.69}$$

The Gershgorin circle theorem can be used to bound the eigenvalue $\lambda_1^{(\Delta)}$ in terms of the matrix elements $\Delta M_{z,z'}$. For a real square matrix A, the Gershgorin circle theorem states that the i'th eigenvalue satisfies the inequality

$$|\lambda_i - A_{ii}| \leq \sum_{j \neq i} |A_{ij}|,$$

which implies that

$$|\lambda_i| \leq \sum_j |A_{ij}| \tag{3.70}$$

Applying Eq. (3.70) to any eigenvalue $\lambda_z^{(\Delta)}$ of ΔM, we have

$$|\lambda_z^{(\Delta)}| \leq \sum_{z'} |\Delta M_{zz'}| \leq Z U_\delta. \tag{3.71}$$

Since Eq. (3.71) only holds if $|\Delta M_{zz'}| \leq U_\delta$ for all z', the probability of the bound is at least $(1-\delta)^Z > 1 - Z\delta$. Combining Eq. (3.71) with Eq. (3.69), we recover Eq. (3.68). □

We now use the element-wise bound on $B_{zz'}^{(a)}$ from Lemma 1.7 to apply Lemma 1.8 to the minimal eigenvalue of the inverse covariance matrix $B^{(a)}$, which immediately translates into an upper bound on the maximal eigenvalue of $(B^{(a)})^{-1}$.

Lemma 1.9. *Under the same conditions as Lemma 1.7, the minimal eigenvalue $\lambda_1^{(a)}(T)$ of the empirical inverse covariance matrix $\frac{1}{T}B^{(a)}(T)$ satisfies the lower bound*

$$\lambda_1^{(a)}(T) \geq \lambda_{\min}^{(a)}(T)/\tilde{c}, \tag{3.72}$$

where $\lambda_{\min}^{(a)}(T)$ is the minimal eigenvalue of $\Sigma^{(a)}(T)$ defined in Eq. (3.52), with probability at least $1 - \delta_\lambda$, where

$$\delta_\lambda := \frac{Z^3}{(\lambda_{\min}^{(a)}(T))^2} \left(\pi_{\min} - \tilde{c}^{-1}\right)^{-2} \frac{2}{T\gamma_\phi}(c + \log\log(1/\rho_{\min})), \tag{3.73}$$

for any $\tilde{c} \in (1, \tilde{c}_{\max})$, *with*

$$\frac{1}{\tilde{c}_{\max}} = \pi_{\min} - \frac{Z}{\lambda_{\min}^{(a)}(T)} \sqrt{\frac{2}{T\gamma_\phi}(c + \log\log(1/\rho_{\min}))}. \tag{3.74}$$

Proof. Recalling Eq. (3.67), we apply Lemma 1.8 with

$$\bar{M} \to \frac{1}{T}\mathbb{E}[B^{(a)}(T)], \quad M \to \frac{1}{T}B^{(a)}(T), \quad U_\delta \to \sqrt{\frac{1}{\delta}\frac{2}{T\gamma_\phi}(c + \log\log(1/\rho_{\min}))},$$

and have

$$\lambda_1^{(a)}(T) \geq \bar{\lambda}_1^{(a)}(T) - ZU_\delta, \tag{3.75}$$

with probability at least $1 - Z\delta$, where $\lambda_1^{(a)}(T)$ and $\bar{\lambda}_1^{(a)}(T)$ are the minimal eigenvalues of $\frac{1}{T}B^{(a)}(T)$ and $\frac{1}{T}\mathbb{E}[B^{(a)}(T)]$, respectively. Using the fact (Lemma 1.5) that $\frac{1}{T}\mathbb{E}[B^{(a)}(T)] \succcurlyeq \pi_{\min}\Sigma^{(a)}(T)$, or equivalently $\frac{1}{T}\mathbb{E}[B^{(a)}(T)] = \pi_{\min}\Sigma^{(a)}(T) + \text{PSD}$ where PSD is a positive semidefinite symmetric matrix with non-negative minimal eigenvalue, and applying the Weyl inequality again (as in Lemma 1.8), we have $\bar{\lambda}_1^{(a)}(T) \geq \pi_{\min}\lambda_{\min}^{(a)}(T)$, and thus,

$$\lambda_1^{(a)}(T) \geq \pi_{\min}\lambda_{\min}^{(a)}(T) - ZU_\delta. \tag{3.76}$$

Defining

$$\tilde{c}^{-1} := \pi_{\min} - \frac{Z}{\lambda_{\min}^{(a)}(T)} \sqrt{\frac{1}{\delta}\frac{2}{T\gamma_\phi}(c + \log\log(1/\rho_{\min}))}, \tag{3.77}$$

Eq. (3.76) takes the form of Eq. (3.72), with $\tilde{c}$ inheriting its range, as stated in Lemma 1.9 above, from the range of $\delta \in (0, 1)$. Inverting Eq. (3.77) to express the probability $\delta_\lambda := Z\delta$ in terms of other parameters, we recover Eq. (3.73). □

3.4.5 Final Bound on Estimator Error

In the preceding sections, we derived high-probability bounds for the empirical covariance matrix $(B^{(a)})^{-1}$ and the error vector $g^{(a)}$. In this section, we combine these results to derive Theorem 1, a high-probability upper bound on the estimator error $\hat{\mu}^{(a)} - \mu_\star^{(a)} = (B^{(a)})^{-1}g^{(a)}$:

Proof of Theorem 1. From Lemma 1.4, we have $(g_z^{(a)}/T)^2 \leq U_\delta^2$ – using U_δ^2 as a shorthand for the right hand side of Eq. (3.31) – with probability at least $1 - \delta$ for any z, and thus

with probability at least $(1-\delta)^Z > 1 - Z\delta$ for all z. Thus, renaming $\delta \to \delta/Z$, the 1-norm of the estimator error is upper bounded with probability at least $1-\delta$:

$$|\hat{\mu}_z^{(a)} - (\mu_\star^{(a)})_z| \leq \sum_{z'} |((B^{(a)})^{-1})_{zz'}| \cdot |g_{z'}^{(a)}| \leq T \cdot U_{\delta/Z} \sum_{z'} \left| ((B^{(a)})^{-1})_{zz'} \right|. \tag{3.78}$$

The sum over elements $|((B^{(a)})^{-1})_{zz'}|$ can be upper bounded in terms of the Frobenius norm $||(B^{(a)})^{-1}||_F$,

$$\sum_{z'} |((B^{(a)})^{-1})_{zz'}| \leq Z \times \max_{z,z'} |((B^{(a)})^{-1})_{zz'}| \leq Z \sqrt{\sum_{z,z'} |((B^{(a)})^{-1})_{zz'}|^2} = Z||(B^{(a)})^{-1}||_F.$$

The singular value decomposition of $(B^{(a)})^{-1}$, which is symmetric and positive semidefinite, can be written $(B^{(a)})^{-1} = \frac{1}{T} U_a \Lambda_a^{-1} U_a^\top$ where U_a is an orthogonal matrix and Λ_a is the diagonal matrix whose nonzero entries are the eigenvalues of $\frac{1}{T}B^{(a)}$. (Recall that the elements of the matrix $B^{(a)}$ increase linearly with T, with $\frac{1}{T}B^{(a)}$ approaching a constant matrix at large T.) The Frobenius norm of a matrix is unchanged under a (left or right) orthogonal transformation, so

$$T \cdot ||(B^{(a)})^{-1}||_F = ||\Lambda_a^{-1}||_F = \sqrt{\sum_z (\lambda_z^{(a)})^{-2}} \leq \frac{\sqrt{Z}}{\lambda_1^{(a)}},$$

where $\lambda_1^{(a)}$ is the minimal eigenvalue (at time T) of $\frac{1}{T}B^{(a)}$. Thus, $T \cdot \sum_{z'} |(B^{(a)})^{-1})_{zz'}| \leq Z^{3/2}/\lambda_1^{(a)}$. Substituting this into Eq. (3.78) above, and recalling Lemma 1.9, we have

$$|\hat{\mu}_z^{(a)} - (\mu_\star^{(a)})_z| \leq \frac{Z^{3/2}\tilde{c}}{\pi_{\min}\lambda_{\min}^{(a)}(T)} U_{\delta/Z}$$

with probability at least

$$(1-\delta)(1-\delta_\lambda) > 1 - \delta - \delta_\lambda.$$

With the definition of δ_λ in Eq. (3.73), recalling that U_δ^2 refers to the upper limit in Eq. (3.31), and setting $\tilde{c} = 2/\pi_{\min}$ for simplicity, we recover Theorem 1 as stated above. □

Note from Eq. (3.73) (with $\tilde{c} = 2$) that in order for the probability of the bound to become positive, the time T (measured in mixing times $1/\gamma_\phi$) must exceed a minimal threshold value,

$$T\gamma_\phi > \frac{8Z^3}{\pi_{\min}\lambda_{\min}^{(a)}}(c + \log\log(1/\rho_{\min})). \tag{3.79}$$

Before this timescale, insufficient data can be gathered to reliably reduce the variance of the estimator. Once $T\gamma_\phi$ exceeds this threshold value, which is parametrically large in the number of latent states Z, the bound becomes nontrivial.

Chapter 4

Latent Linear Thompson Sampling

In this section we describe our proposed algorithm, which (i) uses a prior model $p(x|z;\theta)$ to learn a least-squares estimator for the latent transition matrix, and (ii) uses the learned transition model to build least-squares reward estimators in the linear bandit framework.

4.1 Least-Squares Transition Matrix Estimation

Given the current observation x_t and true parameters $\theta^\star$, we define the negative log-likelihood of observing x_t conditional on latent state z as

$$Y_{t,z} := -\log p(x_t|z;\theta^\star), \tag{4.1}$$

which will be the dependent (target) variables in least-squares estimation of $\phi^\star$. Given prior knowledge of the context distributions $p(x|z)$, the negative log-likelihoods $Y_{t,z}$ are known, deterministic functions of x_t and may thus be treated as observed quantities.

Given data $x_{1:t-1}$, the expected negative log-likelihoods at time t are

$$\begin{aligned}\mathbb{E}[Y_{t,z}|x_{1:t-1}] &= \sum_{z',z''} p^\star(z_{t-1}=z'|x_{1:t-1})\phi^\star_{z',z''}\cdot \mathbb{E}_{x_t\sim p(\cdot|z_t=z'')}[-\log p(x_t|z_t=z)] \\ &= \sum_{z',z''} p^\star_{t-1}(z')\phi^\star_{z',z''}\cdot H_{z'',z}, \end{aligned} \tag{4.2}$$

where the cross-entropy $H_{z'',z'}$ was defined in (3.2). Defining $H \in \mathbb{R}_+^{Z\times Z}$ as the (non-negative) matrix[1] of cross-entropies, Eq. (4.2) can be written in matrix form[2]

$$\mathbb{E}[Y^\top] = P_\star^\top \Phi^\star H, \tag{4.3}$$

where $Y, P_\star \in \mathbb{R}_+^{Z\times T}$ are non-square matrices, and $P^\star$ is the matrix whose columns are posterior probability vectors at *preceding* timesteps, that is $(P_\star)_{z,t} = p^\star_{t-1}(z)$. Defining the linear transformation

$$\Phi^\star_H := \Phi^\star H \tag{4.4}$$

of the transition matrix, we can express Y in the standard linear regression form

$$Y^\top = P_\star^\top \Phi^\star_H + \epsilon^\top, \tag{4.5}$$

where $\Phi^\star_H$ contains the unknown parameters, Y is the dependent variable, $P_\star$ plays the role of the independent variable, and $\epsilon \in \mathbb{R}^{Z\times T}$ is a matrix of residuals, $\epsilon_{t,z} := Y_{t,z} - \mathbb{E}[Y_{t,z}|x_{1:t-1}]$ with zero expectation value. Introducing a model estimate $\hat\Phi_H$ and defining predictions

$$\hat Y_{z,t} = (P_\star^\top \hat\Phi_H)_{t,z}, \tag{4.6}$$

the quadratic loss

$$\sum_{t,z} (Y_{t,z} - \hat Y_{t,z})^2, \tag{4.7}$$

is minimized by the least-squares matrix[3] estimator

$$\hat\Phi_H := (P_\star P_\star^\top + \lambda_\phi \mathbb{1}_Z)^{-1} P_\star Y^\top. \tag{4.8}$$

where $\lambda_\phi > 0$ is a regularization parameter which ensures invertibility[4], and $\mathbb{1}_Z$ is the Z-dimensional identity matrix. We can now use Eq. (4.4) to define a corresponding estimator for the transition matrix, $\hat\Phi(P = P^\star)$, where

$$\hat\Phi(P) := (PP^\top + \lambda_\phi \mathbb{1}_Z)^{-1} P Y^\top H^{-1} \tag{4.9}$$

[1] We will assume throughout that no two distributions $p(x|z)$ and $p(x|z')$ are identical for $z \neq z'$, and consequently that the cross-entropy matrix H is full-rank and invertible.

[2] For any matrix A with latent-state z and timestep dimensions t, we index rows with z and columns with t.

[3] Since the dependent variable Y_t is vector-valued, with components corresponding to different states z indexing the columns of $Y^\top \in \mathbb{R}_+^{T\times Z}$, $\hat\Phi_H$ is a matrix of distinct least-squares estimators in each column.

[4] When $\lambda_\phi = 0$, we recover the ordinary least-squares (OLS) estimator, which is unbiased but may have higher variance.

for any P.

In practice, we lack oracle access to the true posteriors $p_t^\star$ (which depend on $\Phi^\star$) and must use the model posteriors $\hat{p}$ instead, iteratively updating $\hat{\Phi}(\hat{P})$ and $\hat{p}(\hat{\Phi})$ with Eqs. (3.3) and (4.9). This results in bias in both $\hat{\Phi}$ (due to $\hat{p} \neq p^\star$) and $\hat{p}$ (due to $\hat{\Phi} \neq \Phi^\star$). However, we find empirically that $(\hat{p}, \hat{\Phi})$ do converge jointly to $(p^\star, \Phi^\star)$ under certain conditions (see Section 4.3).

Relatedly, we have assumed that $H = H(\theta = \theta^\star)$ is built from the true distributions $p(x|z)$. More generally, we allow a model estimate $\theta \neq \theta^\star$ in Algorithm 2 below, and empirically study robustness to $\theta \neq \theta^\star$ in Section 4.3.

The estimator $\hat{\Phi}_t$ at time t can be updated sequentially as follows. Initializing $B_\phi = \lambda_\phi \mathbb{1}_Z$ and $F_\phi = \mathbf{0}_{Z \times Z}$ at $t = 0$ as matrices which will be used to sequentially update (respectively) the matrices $(PP^\top + \lambda_\phi \mathbb{1}_Z)$ and $PY^\top$ in Eq. (4.9), after each observation x_t we compute $Y_{t,z} = p(x_t|z; \theta^\star)$ for each z, and update

$$B_\phi \leftarrow B_\phi + \hat{p}_{t-1}\hat{p}_{t-1}^\top, \quad F_\phi \leftarrow F_\phi + \hat{p}_{t-1}Y_t^\top \tag{4.10}$$

$$\hat{\Phi}_t = B_\phi^{-1} F_\phi H^{-1} \tag{4.11}$$

where $Y_t, \hat{p}_{t-1} \in \mathbb{R}^Z$ denote column vectors, and at each step we update $\hat{p}$ with Bayes' rule,

$$\hat{p}_t \propto e^{-Y_t} \odot (\hat{\Phi}_t \hat{p}_{t-1}), \tag{4.12}$$

where $e^{Y_t} \in \mathbb{R}^Z$ is the vector with elements $e^{Y_{t,z}}$, and $\odot$ indicates the element-wise vector product. While $\hat{\Phi}$ is unconstrained, and thus may not represent a well-defined matrix of transition probabilities (allowing elements of $\hat{p}$ to likewise be negative), in our experiments we found convergence to the ground truth probabilities.

4.1.1 Special Case: I.I.D. Latent Variables

The case in which the latent state is resampled i.i.d at each timestep can be recovered as the special case in which $\Phi^\star$ is a rank-one matrix with identical columns $\phi^\star$. In this case, at each timestep the random variable x_t is generated from a fixed distribution, and hence the random variables $Y_{t,z}$ are generated from fixed distributions. Eq. (4.2) simplifies to

$$\mathbb{E}[Y_{t,z}] = \sum_{z'} \phi_{z'}^\star \cdot H_{z',z} = ((\phi^\star)^\top H)_z, \tag{4.13}$$

and is independent of the prior history $x_{1:t-1}$. Since each negative log-probability vector $Y_t \in \mathbb{R}^Z$ is an unbiased estimate for the same mean vector, the sample mean $\hat{Y}^{(t)} := t^{-1} \sum_{t'=1}^{t} Y_t$ will converge to the true mean vector $(\phi^\star)^\top H$. The estimator

$$\hat{\phi}_t^\top := \hat{Y}^{(t)} H^{-1} \tag{4.14}$$

is unbiased, since $\mathbb{E}[\hat{\phi}_t^\top] = (\phi^\star)^\top H H^{-1} = (\phi^\star)^\top$.

Compared to the more general non-stationary case, $\hat{\phi}_t$ no longer depends on the history of past model posteriors $\hat{p} \neq p^\star$, but only on the log-likelihoods Y_t and cross-entropy matrix H, quantities obtained from the known distributions $p(x|z;\theta^\star)$. Of course, if the true parameters $\theta^\star$ are not perfectly known, the estimator of latent-state probabilities, Eq. (4.14), will become biased and no longer be guaranteed to converge to $\phi^\star$ as $t \to \infty$.

4.2 Latent Linear Thompson Sampling (L^2TS)

As described in Section 3.2, we treat the model posterior over the current latent state $\hat{p}_t$, Eq. (3.3) – which is updated jointly with the transition estimator $\hat{\Phi}$ – as a context feature vector in the linear bandit setting, $c_t = \hat{p}_t$. Our algorithm, L^2TS, combines[5] the latent posterior ($\hat{p}$) and transition matrix ($\hat{\Phi}$) estimation of Section 4.1 with the ($\hat{p}$-conditioned) reward estimation of LinTS into an end-to-end pipeline for the confounded and non-stationary

[5]The L^2 refers to both "latent linear" and "double least squares"; the latter is in reference to the joint least-squares estimation of the latent transition probabilities and reward parameters.

latent bandit setting of Section 3.1.

Algorithm 2: Linearized Latent Thompson Sampling (L^2TS)

Input:
Prior over latent state, $\hat{p}_0 \in [0,1]^Z$
Model distributions $p(x|z;\theta)$
$F_\phi = \mathbf{0}_{Z\times Z}$, $B_\phi = \lambda_\phi \mathbb{1}_Z$; $\lambda_\phi > 0$
$f_\mu^{(a)} = \mathbf{0}_Z$, $B^{(a)} = \lambda_\mu \mathbb{1}_Z$, for $a \in \mathcal{A}$; $\lambda_\mu > 0$
$\tilde{\sigma}_r^{(a)} > 0$ for $a \in \mathcal{A}$
Precompute $H(\theta)$, from Eqs. (3.2).
for $t \leftarrow 1, 2, \ldots$ **do**
Observe x_t; set $Y_{t,z} = -\log p(x_t|z)$, for $z \in \mathcal{Z}$
Update transition matrix estimate,
$B_\phi \leftarrow B_\phi + \hat{p}_{t-1}\hat{p}_{t-1}^\top$, $F_\phi \leftarrow F_\phi + \hat{p}_{t-1}Y_t^\top$
$\hat{\Phi} = B_\phi^{-1} F_\phi H^{-1}$
Update posterior, $\hat{p}_t \propto e^{-Y_t} \odot (\hat{\Phi}\hat{p}_{t-1})$
Sample $\mu^{(a)} \sim \mathcal{N}(\hat{\mu}^{(a)}, (\tilde{\sigma}_r^{(a)})^2 (B^{(a)})^{-1})$ for $a \in \mathcal{A}$
Select action $a = \arg\max_{a'} \sum_z \hat{p}_t(z) \mu_z^{(a')}$
Observe r_t
Update mean reward estimates:
$B^{(a)} \leftarrow B^{(a)} + \hat{p}_t \hat{p}_t^\top$, $f_\mu^{(a)} \leftarrow f_\mu^{(a)} + \hat{p}_t r_t$
$\hat{\mu}^{(a)} = (B^{(a)})^{-1} f_\mu^{(a)}$

In Algorithm 3 we show a variant of L^2TS for stationary environments in which a non-dynamical latent state is resampled i.i.d. at each round (Section 4.1.1). In this case, Eq. (4.14) is used to estimate the prior probabilities of latent states, instead of the transition

matrix estimate, Eq. (4.11), used by Algorithm 2 for non-stationary environments.

Algorithm 3: Linearized Latent Thompson Sampling: stationary tasks

Input:

Model distributions $p(x|z;\theta)$

$f_\mu^{(a)} = \mathbf{0}_Z$, $B^{(a)} = \lambda_\mu \mathbb{1}_Z$, for $a \in \mathcal{A}$; $\lambda_\mu > 0$

$\tilde{\sigma}_r^{(a)} > 0$ for $a \in \mathcal{A}$

Precompute $H(\theta)$, from Eq. (3.2).

for $t \leftarrow 1, 2, \ldots$ **do**

Observe x_t; set $Y_{t,z} = -\log p(x_t|z)$, for $z \in \mathcal{Z}$

Update sample mean negative log-likelihoods,

$\hat{Y} \leftarrow t^{-1} \sum_{t'=1}^{t} Y_t$

Update estimate of latent-state prior probabilities,

$\hat{\phi} \leftarrow \hat{Y} H^{-1}$

Compute posterior, $\hat{p}_t \propto e^{-Y_t} \odot \hat{\phi}$

Sample $\mu^{(a)} \sim \mathcal{N}(\hat{\mu}^{(a)}, (\tilde{\sigma}_r^{(a)})^2 (B^{(a)})^{-1})$ for $a \in \mathcal{A}$

Select action $a = \arg\max_{a'} \sum_z \hat{p}_t(z) \mu_z^{(a')}$

Observe r_t

Update mean reward estimates:

$B^{(a)} \leftarrow B^{(a)} + \hat{p}_t \hat{p}_t^\top, \quad f_\mu^{(a)} \leftarrow f_\mu^{(a)} + \hat{p}_t r_t$

$\hat{\mu}^{(a)} = (B^{(a)})^{-1} f_\mu^{(a)}$

4.3 Experiments

In this section, we conduct experiments to compare L^{2}TS (Algorithm 2) with (i) the theoretical regret bound in a setting with Gaussian data and a non-dynamical latent state, and (ii) relevant baselines in a setting with discrete variables and non-stationarity from a dynamical latent state. We describe the details of these experiments in Sections 4.3.1 and 4.3.2 before presenting results.

4.3.1 Stationary & Gaussian Task: Regret Scaling

Numerical Details: Environment We consider an environment with $Z = 2$ latent states, $K = 2$ actions, and Gaussian context and reward distributions. At each timestep the (non-dynamical) latent state is resampled i.i.d., with prior probabilities $\phi_z^\star$ set to $(\phi_0^\star, \phi_1^\star) =$

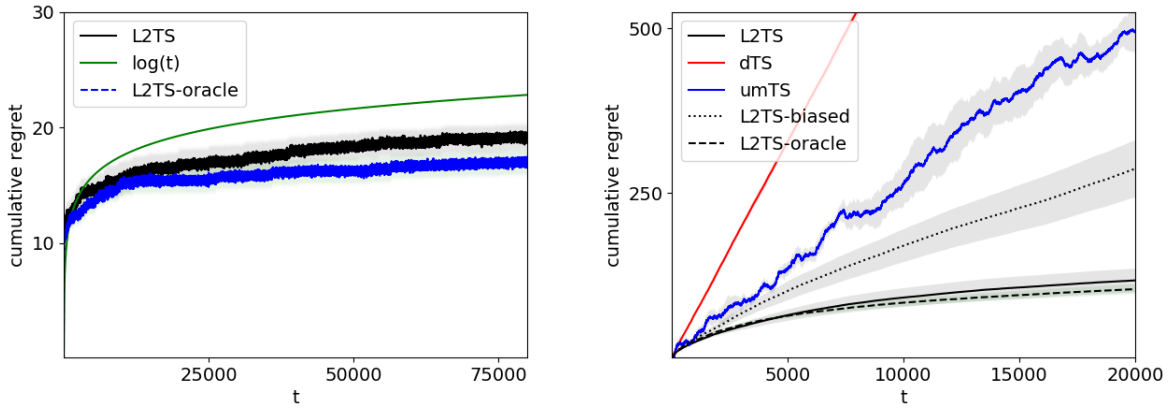

Figure 4.1: **Left:** Mean cumulative regret of L^{2}TS (Algorithm 3) in the stationary Gaussian environment of Section 4.3.1, compared to the optimal scaling $\log(t) + const.$, and to an oracle policy which knows the true latent transition model. **Right:** Mean cumulative regret of L^{2}TS (Algorithm 2) and baseline algorithms in the non-stationary discrete-variable setting of Section 4.3.2, with shaded regions indicating variance over 10 episodes (4 for umTS). L^{2}TS strongly outperforms baselines, approximates the oracle policy, and degrades gracefully when limited by biased models $p(x|z)$.

$(0.3, 0.7)$. The ground truth context distributions $p(x|z)$ are set to Gaussian distributions with mean values $(\mu_{z=0}, \mu_{z=1}) = (0, 1)$ and standard deviations $(\sigma_{z=0}, \sigma_{z=1}) = (0.1, 0.2)$. The reward distributions $p(r|z, a)$ are also Gaussian distributions, with mean values $\mu_z^{(a)}$ set to $(\mu_0^{(0)}, \mu_0^{(1)}, \mu_1^{(0)}, \mu_1^{(1)}) = (1, 2, 2, 1)$, and a shared standard deviation $\sigma_z^{(a)} = 1$ for all (z, a).

Numerical Details: Algorithm We used Algorithm 3, the variant of L^{2}TS for stationary settings, with covariance matrix hyperparameters $\lambda_\mu = \tilde{\sigma}_r^{(a)} = 1$. (We observed comparable performance with the more general Algorithm 2, which handles both stationary and non-stationary settings.) We also compared to an oracle version of L^{2}TS which accessed the ground truth probability vector $\phi^\star$.

4.3.2 Non-Stationary Task with Discrete Variables

Numerical Details: Environment We consider an environment with $Z = 3$ latent states, $K = 2$ actions, and integer-valued context $x \in \{1, 2, 3, 4\}$. The ground-truth

context distributions are set to

$$p(x|z) = \begin{bmatrix} 0.05 & 0.85 & 0.09 & 0.01 \\ 0.01 & 0.19 & 0.79 & 0.01 \\ 0.01 & 0.01 & 0.59 & 0.39 \end{bmatrix}, \tag{4.15}$$

such that x contains significant information about z. The ground-truth reward distributions,

$$p^{\intercal}(r|a=1,z) = \begin{bmatrix} 0.2 & 0.5 & 0.8 \end{bmatrix}; p^{\intercal}(r|a=2,z) = \begin{bmatrix} 0.3 & 0.4 & 0.5 \end{bmatrix}, \tag{4.16}$$

contain somewhat less information for distinguishing between latent states. We choose a transition matrix over latent states that favors the diagonal, intended as a toy model of real-world settings where latent state transitions are predominantly local in latent space:

$$p(z_t|z_{t-1}) = \begin{bmatrix} 0.7 & 0.25 & 0.05 \\ 0.25 & 0.5 & 0.25 \\ 0.05 & 0.25 & 0.7 \end{bmatrix}. \tag{4.17}$$

Numerical Details: Algorithm We used L^{2}TS (Algorithm 2) with covariance matrix hyperparameters $\lambda_\phi = \lambda_\mu = \tilde{\sigma}_r^{(a)} = 1$.

Baselines. We compare with two strong baselines: (1) Uncertain Model Thompson Sampling (umTS) [32]: We adapt umTS, which uses particle filtering to maintain a posterior over reward models and latent transition matrices, to our setting by using prior knowledge of $p(x_t|z;\theta^\star)$ for posterior updates. (In the original graphical setting of [32], the latent state only influences rewards, and not contexts.) As such, umTS accesses the same prior knowledge as L^{2}TS, and also uses reward data for latent state inference. In contrast, L^{2}TS does not assume a parameteric reward model, and relies exclusively on context data to learn about the latent space. (2) Discounted Thompson Sampling (dTS) [47]: We extend dTS to maintain success ($r = 1$) and failure ($r = 0$) counts for each context-action pair (x, a), and allow dTS to use the true dynamics timescale to set the discount factor γ. This gives it the advantage of partial prior knowledge of the hidden Markov dynamics, but not of the context likelihood as with L^{2}TS or umTS.

In addition to baseline algorithms, we also consider (i) an oracle variant of L^{2}TS which uses the true posterior $p_t^\star(z) := p_t(z|x_{1:t};\phi^\star,\theta^\star)$ to learn reward parameters, and (ii) a biased variant of L^{2}TS which uses $N_z = 10$ offline samples $x \sim p(x|z)$ for each z to construct an approximate model $\theta \approx \theta^\star$ based on counts of sampled discrete values x.

Baseline Details. We allowed the discounted Thompson Sampling (dTS) algorithm to access the true transition matrix by setting its discount factor to $\gamma = Z^{-1} \sum_z \phi^\star_{z,z}$. For umTS [32], we used $N = 1000$ particles with a minimum effective sample size $ESS_{min} = 50$ for particle resampling.

4.3.3 Results

In Figure 4.1 (Left), we show the asymptotic regret of L^{2}TS in the stationary Gaussian environment (Section 4.3.1), compared to (i) the optimal logarithmic scaling, and to (ii) an oracle version of L^{2}TS which uses the true latent probabilities $\phi^\star$ and oracle access to the true reward model $p(r|z, a)$ to compute latent-state posteriors. L^{2}TS performs well asymptotically, learning the true latent-state probabilities and approaching the performance of the oracle.

Figure 4.1 (Right) shows the cumulative regret[6], averaged over 10 episodes, for all algorithms (except umTS, for which we average over 4 episodes). L^{2}TS significantly outperforms the baseline algorithms. While umTS models the true latent structure and context distributions $p(x|z)$, it suffers from asymptotically linear regret as a result of failure of its particle-based posterior to converge.[7] (We found that umTS was able to learn the true transition matrix with $\simeq 10\%$ accuracy, and mean rewards somewhat less accurately.) Discounted TS performs most poorly due to its inability to model the latent space or to transfer information gained across different discrete contexts. L^{2}TS closely approximates the oracle, indicating asymptotic convergence of the learned transition model. While the biased variant suffers from linear regret, since θ is never improved, L^{2}TS degrades gracefully in light of the limitation – in this case to only $N_z = 10$ samples $x \sim p(\cdot|z)$ for each z.

[6] Regret is defined relative to the optimal policy conditioned on the true task parameters, which selects at each timestep the action a maximising $(p^\star_t)^\top \mu^{(a)}_\star$.

[7] We expect this performance to improve with stronger particle resampling methods which resample by interpolating between existing particles, or update particle positions as well as weights [18].

Chapter 5

Slow Dynamics and Latent State Distinguishability

As noted below Lemma 1.3, our derivation of Theorem 1 was conservative in that it did not make use of the context distribution parameters $\theta^\star$ – which capture the amount of information about the latent state contained in context observations – to strengthen the bound. In this section, we derive for a simpler algorithm an instantaneous regret bound which captures this effect.

We introduce Algorithm 4 (mw-zTS), a naive and limited variation on Algorithm 2 (L^{2}TS) which uses fixed estimate reward parameters and a memory window τ as a rough proxy for the latent dynamics timescale[1], in order to make theoretical analysis more tractable. Algorithm 4 Thompson samples latent states z instead of reward parameters. When Algorithm 4 accesses the true reward parameters, $\hat{\mu}^{(a)} = \mu_\star^{(a)}$, its performance is limited exclusively by uncertainty or error in the model posterior $\hat{p}_t$. As such, it can be

[1] A memory window can be a useful alternative to a full transition model Φ when the latent state changes rarely, and when learning Φ is difficult, e.g. due to large Z.

used as a theoretical tool to study the influence of $p(x|z)$ on inference of the latent state.

Algorithm 4: Memory Window Latent-State Sampling (mw-zTS)

Input: Memory window $\tau \in \mathbb{N}$
Model distributions $p(x|z,\theta)$
Reward estimates $\hat{\mu}^{(a)} \in \mathbb{R}^Z$, for $a \in \mathcal{A}$
for $t \leftarrow 1, 2, \ldots$ **do**
Observe x_t and update posterior:
$\hat{p}_t(z) \propto \prod_{t'=\max(1,t-\tau)}^{t} p(x_{t'}|z;\theta)$
Sample $z \sim \hat{p}_t(\cdot)$; select action $a = \arg\max_{a'} \hat{\mu}_z^{(a')}$
Observe r_t

Theorem 2 below uses an alternate definition of the instantaneous regret incurred at time t,

$$\delta\mathcal{R}_\pi^{(\text{oracle})}(t) := \mathbb{E}_{\text{oracle}}[r_t] - \mathbb{E}_\pi[r_t], \tag{5.1}$$

relative to an oracle policy which accesses the true latent state $z_t^\star$ and earns a mean reward,

$$\mathbb{E}_{\text{oracle}}[r_t] = \max_a((\mu_\star^{(a)})_{z_t^\star}). \tag{5.2}$$

Here, the oracle may be viewed as a version of Algorithm 4 which uses an oracle posterior, $\hat{p}_t(z) \to \mathbf{1}(z = z_t^\star)$, along with the true reward parameters. As such, this definition of regret isolates the effect of probability mass $\hat{p}_t(z)$ placed on states $z \neq z_t^\star$.

Before stating our result, we also introduce the pair-wise KL divergences between conditional context distributions,

$$D_{z^\star,z}(\theta) := \mathbb{E}_{x\sim p(\cdot|z^\star;\theta)}\left[\log(p(x|z^\star;\theta)/p(x|z;\theta))\right], \tag{5.3}$$

which quantify the distinguishability of a given pair of latent states $(z^\star, z)$ through observations $x \sim p(\cdot|z^\star)$, and will be used to bound the probabilities $\hat{p}_t(z)$ assigned to $z \neq z_t^\star$.

With these definitions, the regret at a single timestep incurred by mw-zTS satisfies a high-probability upper bound:

Theorem 2. *Given maximal state change probability* $1/L^\star = max_z(1 - \phi_{z,z}^\star)$, *reward gaps* $\Delta_{z^\star,z} := \mu_{z^\star}^{(a(z^\star))} - \mu_{z^\star}^{(a(z))}$ *where* $a(z) = \arg\max_{a'} \mu_z^{(a')}$ *with* $\{\mu^{(a)}\} = \{\mu_\star^{(a)}\}$, *a measure* $\rho_t := (\Phi^\star)^t\rho_0$ *over latent states, pairwise KL divergences* $D_{z^\star,z}(\theta^\star)$, *and log-likelihood variances*

$$\sigma^2_{z^\star,z}(\theta) := \mathbb{E}_{x\sim p(\cdot|z^\star,\theta)}[\log^2(p(x|z^\star;\theta)/p(x|z;\theta))] - D^2_{z^\star,z}(\theta), \tag{5.4}$$

the instantaneous regret of Algorithm 4 relative to the oracle policy which observes the latent state satisfies

$$\delta\mathcal{R}^{(\text{oracle})}_{\text{mw}-z\text{TS}}(t) \leq \frac{\tau}{L^\star}\Delta_{\max} + \sum_{z^\star,z} \rho_t(z^\star) e^{-\tau D_{z^\star,z} + \sqrt{\frac{\tau}{\delta}}\sigma_{z^\star,z}} \Delta_{z^\star,z} \tag{5.5}$$

where $\Delta_{\max} := \max_{z,z'} \Delta_{z,z'}$*, with probability at least* $1 - Z^2\delta$*, for any* δ *such that* $\tau D_{z^\star,z} - \sqrt{\tau/\delta}\sigma_{z^\star,z} > 0$ *for all* $(z^\star, z)$*.*

Proof (Outline). The complete proof (see Section 5.1) involves (i) bounding the probability of histories in which the latent state has changed within the past memory window, which scales as $\tau/L^\star$, and (ii) following the method of [27] to upper bound the posterior probability $p_t(z \neq z^\star)$ assigned to latent states other than the true state, which is controlled by the KL divergences $D_{z^\star,z}(\theta)$ of Eq. (5.3),

$$\mathbb{E}_{x_{t-\tau:t}\sim p(\cdot|z^\star)}\left[p_t(z|x_{t-\tau:t})\right] \propto e^{-\tau D_{z^\star,z}}, \tag{5.6}$$

up to a contribution from the sample variance of observations which can be bounded with high probability for a large memory window τ. □

Theorem 2 expresses the joint influence of the timescale of non-stationarity $L^\star$ and the distributional structure of the likelihoods $p(x|z)$ on the agent's knowledge of the evolving latent state. The first term in Eq. (5.5) bounds the regret incurred when the latent state has changed in the previous τ timesteps, while the second term, obtained from an upper bound on posterior probabilities for $z \neq z_t^\star$, Eq. (5.18), bounds the regret incurred due to limited information about the latent state being available in the previous τ timesteps. For a ground-truth state $z^\star$, the posterior probability of another state z decays exponentially in time, with the KL divergence $D_{z^\star,z}$, acting as an average decay rate. (The additional term in the exponent acts as a probabilistic lower bound on the actual decay rate due to variance around the mean, with ϵ controlling the degree of tolerance of rare trajectories for which $z^\star$ is unusually difficult to distinguish from z.)

Since mw-zTS makes the naive assumption that the latent state has not changed in the past τ steps, it can err when τ is too large relative to the dynamical timescale $L^\star$ (the first term in Eq. (5.5)), or when τ is too small, and not enough data are used to distinguish z (the second term). In Section 5.1.3, we show that – in a simplified setting where all states are equally distinguishable – $D_{z^\star,z} = D$ for all $(z^\star, z \neq z^\star)$ – the optimal window scales as $\tau \sim \log(L^\star D)/D$, with regret scaling as $1/L^\star D$, such that when the KL divergences $D_{z^\star,z}$ are large compared to the dynamics timescale, the latent state may be inferred before it

changes. In this regime, the model posterior $\hat{p}_t$ will assign high probability to the true latent state $z_t^\star$, yielding low regret from sampling $z \neq z_t^\star$. We expect similar dependence on $D_{z^\star,z}$ in the estimator error, Eq. (3.11), via the covariance matrices, Eq. (3.13), whose spectra describe the uncertainty in posterior beliefs $p_t^\star$. Theorem 2 pedagogically illustrates the effect of context distributions $p(x|z)$ on the simplified algorithm mw-zTS.

While we've assumed for simplicity that only the context data is used to update the posterior $p(z_t = z|x_{1:t})$, it should be straightforward to extend Theorem 2 to the case of a reward-conditioned posterior $p(z_t = z|x_{1:t}, r_{1:t})$, which will introduce an analogous dependence on KL divergences between reward distributions $p(r|z, a)$ conditioned on different latent states.

5.1 Derivation of Theorem 2

We will define the per-timestep regret as

$$\delta\mathcal{R}_t = \mathbb{E}_{\text{oracle}}[r_t] - \mathbb{E}_{\text{mw}-z\text{TS}}[r_t], \tag{5.7}$$

where the expectation is over latent state histories $z_{1:t}^\star$ (conditioned on $\phi^\star$), sequences of observed contexts $x_{1:t}$ (conditioned on $\theta^\star$), and actions generated by mw-zTS (Algorithm 4) or by the oracle policy which uses the true latent state $z^\star$ to select $a^\star = \arg\max_a (\mu_\star^{(a)})_{z^\star}$. For convenience, we also recall the reward gaps defined in Theorem 2,

$$\Delta_{z^\star,z} := \mu_{z^\star}^{(a(z^\star))} - \mu_{z^\star}^{(a(z))},$$

where $a(z) := \arg\max_{a'} (\mu_\star^{(a')})_z$, and the measure $\rho_t := (\Phi^\star)^t \rho_0$ over latent states at time t.

5.1.1 Bound on non-stationarity in the environment

We assume, as stated in Theorem 2, that the true transition matrix satisfies a constraint on the state change probabilities, $1 - \phi_{z,z}^\star \leq 1/L^\star$ for all z. Consequently, over a period of $\tau < L^\star$ timesteps, the probability of a change in the latent state is $1 - (1 - 1/L^\star)^\tau < \tau/L^\star$. That is, for $t > \tau$,

$$\begin{aligned}\sum_{z_{1:t}} p(z_{1:t}) &= \sum_{z_{1:t}} \left[\mathbf{1}(z_{t-\tau:t} = z_t) + (1 - \mathbf{1}(z_{t-\tau:t} = z_t))\right] p(z_{1:t}) \\ &\leq \frac{\tau}{L^\star} + \sum_{z_{1:t}} \mathbf{1}(z_{t-\tau:t} = z_t) p(z_{1:t}).\end{aligned}$$

This allows us to express the regret in terms of histories without a latent state change in the most recent τ steps.

$$\delta\mathcal{R}_t \leq \frac{\tau}{L^\star}\Delta_{\max} + \sum_{z^\star,z} p_{t,\tau}(z^\star|z)\Delta_{z^\star,z}, \tag{5.8}$$

where we've used the fact that $\Delta_{z_t,z'} \leq \Delta_{\max} := \max_{z,z'} \Delta_{z,z'}$, and the normalization,

$$\sum_{x_{1:t}} p(x_{1:t}|z_{1:t}) \sum_{z'} p(z_t = z'|x_{1:t}) = 1,$$

to obtain the first term, and where

$$p_{t,\tau}(z|z^\star) \propto \sum_{z^\star_{1:t}} \mathbf{1}(z^\star_{t-\tau:t} = z^\star) \sum_{x_{1:t}} p(x_{1:t}|z^\star_{1:t}) p(z_t = z|x_{1:t}) \tag{5.9}$$

is the posterior over the latent state at time t, averaged over observed and ground-truth latent sequences for which the true latent state $z^\star$ has not changed in the past τ timesteps. Eq. (5.8) allows us to bound the regret in terms of an average over histories for which the latent state has not changed recently, up to an error term that is small when $\tau \ll L^\star$.

5.1.2 Bound on the posterior

We now derive a bound on the posterior probability assigned by Algorithm 4 to latent states $z \neq z^\star$ differing from the true state. To simplify notation, in this section we define

$$\mathbb{E}_z[f(x_{t_1:t_2})] := \mathbb{E}_{x_{t_1},\ldots,x_{t_2}\sim p(\cdot|z)}[f(x_{t_1:t_2})] \tag{5.10}$$

as the expectation of any function of the sequence $x_{t':t}$, over sequences with $x_t \sim p(\cdot|z)$ for $t_1 \leq t \leq t_2$.

We would like to bound the averaged posterior $p_{t,\tau}(z'|z)$ in Eq. (5.8) in terms of $L^\star$ and the known observation likelihood $p(x|z)$. Recalling that Algorithm 4 uses a posterior proportional to the likelihood of the most recent τ observations, we have

$$\begin{aligned} p_{t,\tau}(z|z^\star) &:= \mathbb{E}_{z^\star}\left[\frac{\prod_{t'=t-\tau}^t p(x_{t'}|z)}{\sum_{z'}\prod_{t'=t-\tau}^t p(x_{t'}|z')}\right] \\ &= \mathbb{E}_{z^\star}\left[\frac{e^{\ell_z}}{\sum_{z'} e^{\ell_{z'}}}\right], , \end{aligned} \tag{5.11}$$

which is the expectation over partial sequences $x_{t-\tau:t}$ of a softmax distribution with logits given by sums of likelihoods

$$\ell_z := \sum_{t'=t-\tau}^{t} \log p(x_{t'}|z), \tag{5.12}$$

which have expectation value

$$\mathbb{E}_{z^\star}[\ell_z] = \tau \times \mathbb{E}_{x\sim p(\cdot|z^\star)}[\log p(x|z)] = -\tau(D_{z^\star,z} + H_{z^\star}). \tag{5.13}$$

Following [27], we will bound the posterior in terms of the KL divergences $D_{z^\star,z}$, which control the posterior probabilities for states z when $z^\star$ is the true state, as well as the variance of log-probabilities ℓ_z, which determines the variation of the posterior across histories. The softmax probabilities can be upper bounded in terms of the difference of log-probabilities for z and $z^\star$. For $z \neq z^\star$,

$$\frac{e^{\ell_z}}{\sum_{z'} e^{\ell_{z'}}} = \frac{e^{\ell_z}}{e^{\ell_{z^\star}} + e^{\ell_z} + \sum_{z'\neq z,z^\star} e^{\ell_{z'}}} \leq e^{-(\ell_{z^\star} - \ell_z)}. \tag{5.14}$$

The exponent is a random variable (determined by $x_{t-\tau:t}$) with mean $-\tau D_{z^\star,z}$ and it is straightforward to show that its variance is $\tau\sigma^2_{z^\star,z}$, where

$$\sigma^2_{z^\star,z} := \mathbb{E}_{x\sim p(\cdot|z^\star)}\left[\log^2\left(\frac{p(x|z^\star)}{p(x|z)}\right)\right] - D^2_{z^\star,z} \tag{5.15}$$

We can bound its deviation from the mean with high probability using Chebyshev's inequality, which states that for a random variable X with variance $0 < \sigma_X^2 < \infty$,

$$P(|X - \mathbb{E}[X]| \geq \sigma_X/\sqrt{\epsilon}) \leq \epsilon. \tag{5.16}$$

In our case, this is

$$P(|\ell_{z^\star} - \ell_z - \tau D_{z^\star,z}| \geq \sigma_{z^\star,z}\sqrt{\tau/\epsilon}) \leq \epsilon. \tag{5.17}$$

Thus, removing the expectation value over $x_{t-\tau:t}$ in Eq. (5.11), we have

$$p_{t,\tau}(z|z^\star) \leq \exp\left[-\tau D_{z^\star,z} + \sqrt{\tau/\epsilon}\sigma_{z^\star,z}\right] \tag{5.18}$$

with probability $1-\epsilon$.

Using Eq. (5.18) to bound $p_{t,\tau}(z'|z)$ in Eq. (5.8), we arrive at the final regret bound, Eq. (5.5) in Theorem 2, which holds with reduced probability of at least $1 - Z^2\epsilon$ because we conservatively require the Chebyshev bound for each of the $Z(Z-1) < Z^2$ pairs $(z^\star, z \neq z^\star)$.[2]

[2]It may be possible to improve this scaling by exploiting relationships between the distributions $p(x|z)$ for different z.

5.1.3 Optimized Regret Bound for Uniform Model

For generic values of $D_{z,z'}$, $\sigma_{z,z'}$, and $\Delta_{z,z'}$, minimizing Eq. (5.5) with respect to τ is complicated. To better understand the bound with optimal τ, we consider the simplified case where the observation and reward parameters (θ, μ) are fixed to values such that the matrices $D_{z,z'}(\theta^\star)$, $\sigma_{z,z'}(\theta^\star)$, and $\Delta_{z,z'}(\mu^\star)$ are uniform (except for being zero on the diagonal):

$$D_{z,z'} = D \cdot \delta_{z \neq z'}, \;\; \sigma_{z,z'} = \sigma \cdot \delta_{z \neq z'}, \;\; \Delta_{z,z'} = \Delta \cdot \delta_{z \neq z'}. \tag{5.19}$$

Although this is a simplified model, it suffices to illustrate the roles of the parameters in the more general case, via the following result.

Corollary 2.1. *In the restricted case where model parameters satisfy the uniform matrix condition, Eq. (5.19), the upper bound in Eq. (5.5) is minimized for*

$$\tau_{\text{opt}}(\delta) = \frac{\log(\delta L^\star D Z)}{\delta D}, \tag{5.20}$$

and takes the form

$$\delta \mathcal{R}_t \leq \Delta \frac{\log(L^\star D Z) + 1}{\delta L^\star D}, \tag{5.21}$$

holding with probability

$$1 - p_{\text{fail}} = 1 - \frac{\delta}{(1-\delta)^2} \frac{\sigma^2}{D} \frac{Z^2}{\log(\delta L^\star D Z)}, \tag{5.22}$$

for any $\delta \in (\delta_{\min}, \delta_{\max})$, where $\delta_{\min} = 1/L^\star D Z$ and $\delta_{\max} < 1$ is determined by setting Eq. (5.22) to zero.

Proof. For the simplified model of Eq. (5.19), Eq. (5.5) takes the form

$$\delta \mathcal{R}_t \leq \Delta \left(\frac{\tau}{L^\star} + Z e^{-\tau D + \sqrt{\tau/\epsilon}\sigma} \right). \tag{5.23}$$

The optimal window depends on the desired confidence parameter ϵ for the bound. Defining $\delta \in (0, 1)$,

$$\delta := 1 - \sigma / D\sqrt{\epsilon \tau}, \tag{5.24}$$

as a parameter controlling the decay rate in the exponent, $-\delta\tau D$, the upper bound can be minimized with respect to τ, leading to the optimal window parameter, Eq. (5.20), for which Eq. (5.23) reduces to the optimal bound, Eq. (5.21), where we've discarded a negative log term for simplicity. Finally, recalling that the bound holds with probability $1 - \epsilon Z^2$, solving Eq. (5.24) for $\epsilon = \sigma^2 / \tau D^2 (1-\delta)^2$, and fixing τ to τ_{opt}, we arrive at Eq. (5.22). □

We see that the upper bound scales as $\frac{1}{L^\star D}$, becoming strong in the limit where latent states can be easily distinguished within the timescale on which they change. Similarly, in the limit of very large D, the probability of the bound, Eq. (5.22), can be chosen very close to one while still allowing for a low regret, $\sim \frac{\Delta}{\delta L^\star D}$. The optimal window τ_{opt} is controlled predominantly by the distinguishability of latent states D, growing only logarithmically with the timescale $L^\star$ of nonstationarity (an indication of the biasing of old data due to possible latent state changes). Furthermore, τ_{opt} increases as δ is decreased, reflecting the fact that higher confidence in the bound requires including more data to reduce sample variance in the empirical log-probabilities.

Chapter 6

Conclusion

In this thesis, we have studied a non-stationary contextual multi-armed bandit problem in which a discrete latent state evolving under Markovian transition dynamics influences observed contexts and rewards. We have:

1. Shown how the nonlinear relationship between observed contexts and rewards (arising from their influence by the same latent state) can be reduced to a linear relationship via Bayesian inference of the latent variable, reducing the problem to the linear bandit setting.

2. Applied a linear bandit estimator for reward parameters, and derived a high-probability upper bound on the error which applies (i) when accurate posterior beliefs over the latent state can be obtained from observed contexts and (ii) after a time horizon which is long compared to the minimal mixing rate of the latent state transition matrix.

3. Introduced a linear regression method for online learning of the latent-space transition matrix, which exploits prior knowledge of conditional distributions $p(x|z)$ for context observations x, such as relative entropies.[1]

4. Introduced a novel algorithm, motivated by the reduction to the linear bandit setting (point 1 above), which (i) jointly learns a model transition matrix and posterior prob-

[1] Like expectation maximization (EM) [9], our method iterates between two steps (updating transition matrix and latent posterior estimates). Unlike EM, our method does not maintain the history $x_{1:t}$ in memory, and thus has an advantage of computational efficiency (at the expense of sample efficiency gained by using the full history $x_{1:t}$) and may be preferred in memory-restricted or compute-limited applications.

abilities for the current latent state, and (ii) uses the latter as input for a Thompson sampling algorithm for linear bandit problems.

5. Derived a regret bound for a related algorithm which reveals the influence on performance of both the latent dynamics and the structure of the conditional context distributions $p(x|z)$ (which determine how informative contexts $x \sim p(x|z)$ are about the current latent state).

Results (1)-(4) above made use of linearity with respect to probabilities of an unobserved latent variable in two ways:

- The linear relationship between latent state probabilities, latent state dependent rewards, and unknown reward parameters.
- The linear relationship between latent state probabilities, latent state dependent contexts, and unknown latent state transition probabilities.

In each case, we used least squares estimation to estimate the unknown parameters ($\Phi^\star$ or $\mu^\star$), using posterior beliefs as dependent variables in the linear regression.

While Algorithm 2 uses a specific method to learn the latent transition matrix which assumes prior knowledge of distributions $p(x|z)$, the high-level approach of treating a posterior belief over latent space as context information is much more general, and can be applied with any method for learning a latent transition model, as well as other linear bandit algorithms. In section 6.1 below, we outline several directions for extending this approach to more general problem settings.

6.1 Directions for Future Research

We have relied on the graphical and functional simplicity of our problem setting, as well as assumptions of prior knowledge, in order to render the theoretical analysis more tractable. Many interesting directions exist for generalizing the problem setting, algorithmic methods, and/or theoretical analysis:

Theoretical Guarantees.

Problem-Dependent Regret Bound. When used in a Thompson sampling algorithm such as L^{2}TS, the estimators $\hat{\mu}^{(a)}$ whose error is upper bounded in Theorem 1 are the mean values of multivariate Gaussian posteriors. As the error in these estimates converges to zero, and

the posteriors concentrate around their mean values, Thompson sampling will converge towards sampling of the true parameters $\mu_\star^{(a)}$. We thus expect that a straightforward extension of Theorem 1 for Thompson sampling will yield a high-probability problem-dependent regret bound.

Transition Matrix Estimation. While Algorithm 3 is guaranteed to converge to the true latent state probabilities, we have not proven that the more general transition matrix estimation of Algorithm 2 is guaranteed to converge to the true transition probabilities. Such a guarantee would make it possible to extend linear bandit estimation error bounds such as Theorem 1, or any resulting regret bound, to the case where the parameters $(\theta^\star, \Phi^\star)$ are unknown. Alternatively, our method can be replaced with any method for online learning of the parameters $(\theta^\star, \Phi^\star)$ of a hidden Markov model, with the corresponding posterior beliefs $\hat{p}_t$ over the latent state again being used as linear bandit context vectors. A convergence guarantee for such a method could be used to bound the error $p_t^\star - \hat{p}_t$ in the estimated posteriors, which could in turn be used to generalize Theorem 1 to the case of fully unknown parameters.

Algorithmic Improvements.

Parameter Uncertainty. When the model posterior probabilities $\hat{p}_t$ differ from the true probabilities $p_t^\star$ (see Section 3.1.1), the linear bandit Thompson sampler in Algorithm 2 is effectively supplied with biased and/or noisy versions of the linear bandit context c_t (as in Algorithm 1). Some recent work [61] has developed linear bandit algorithms with robustness to imperfect observation of context vectors, and could be applied to our case, in which an imperfect estimate $p_t^\star$ plays the role of a noisy or corrupted context vector. A complementary approach to estimator error would be to use an uncertainty estimate for the estimator $\hat{\Phi}$, such as the covariance matrix B_ϕ, to estimate the error or uncertainty in the model posterior $\hat{p}_t$, and make a corresponding correction to increase the entropy of $\hat{p}_t$. More generally, any online Bayesian learning method which maintains an approximate joint posterior $p(z_t, \Phi, \theta)$ could be applied, with the marginal posterior $p(z_t) = \int_{\Phi,\theta} p(z_t, \Phi, \theta)$ being used as a linear bandit context vector.

Inference with Reward Data. Our method uses only context data, and not reward data, for inference of the latent state. It is therefore most useful in settings where context data contain more information for inference of latent variables. One reason for expecting such settings to be generic is that reward is a scalar variable, whereas context data may be a high-dimensional vector, and may thus contain much more information for inference. Similarly, in more complex graphical settings, latent variables may influence multiple observed contexts, but only one reward node. Furthermore, in stationary bandit settings, the latent state is reset after each reward, so information in reward data cannot be used to improve

future actions. In non-stationary settings where rewards contain useful information for latent state inference, L2TS could be easily extended to include a Bayes update of latent state beliefs using reward data, under the assumption of a Gaussian reward likelihood along with the multivariate normal posterior used for Thompson sampling. When mapped to a linear bandit problem, this introduces a dependence of the current context c_t on past rewards $r_{t'<t}$.

Learned Context Distributions. Arguably the most significant limitation of Algorithm 2 is its reliance on prior knowledge of the conditional context distributions $p(x|z;\theta^\star)$. Our method could be extended in various ways to handle cases where these distributions are unknown, or are imperfectly estimated (e.g. from available offline data). For example, the method of Section 4.1 for iteratively updating estimates $\hat{p}$ and $\hat{\Phi}$ for the latent state probability vector and transition matrix could be extended to include an update to a model estimate $\hat{\theta}$, e.g. by gradient ascent on the marginal likelihood, $\sum_z \hat{p}(z)p(x|z;\hat{\theta})$. (Such a method would again be complementary to expectation maximization, trading off accuracy or sample efficiency to reduce requirements for compute and memory.)

Problem Setting Generalizations.

The graphical model of Figure 3.1 could be generalized in various ways:

- The observation x_t could directly influence the reward r_t, and/or an additional observation variable $\tilde{x}_t$ could influence reward, without any correlation with the latent state z_t (e.g. as in [32]). In this case a more general reward model would be needed, but the same strategy of using linearity with respect to latent state probabilities could be followed.

- More complex graphs could also be considered, as in causal bandits literature [37, 60]. In this case, posterior probabilities over any variables which are causal parent or ancestor nodes of the reward node in a directed acyclic graph could be treated as context features for a linear bandit algorithm. Any method for inference of latent variables (including parent or ancestor nodes of the reward) could then be composed with a linear bandit algorithm, following the general structure of L^{2}TS.

- With an additional directed edge from the action a_t to latent state z_{t+1}, Figure 3.1 describes a partially observable Markov decision process (POMDP). In this setting, it would be interesting to explore Thompson sampling as well as more general Bayesian model-based reinforcement learning approaches which use posterior probabilities $p_t(z)$ over the latent state as inputs to a policy and/or value function, treating the posterior as a belief state. Linearity with respect to posterior probabilities could again be used to inform algorithms and theoretical analysis.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 176–184, 2017.

[3] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013.

[4] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the International Conference on Machine Learning*, pages 127–135, 2013.

[5] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, pages 397–422, 2002.

[6] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77, January 2003.

[7] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[8] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

[9] Leonard E. Baum, Ted Petrie, George W. Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[10] Matthew Beal, Zoubin Ghahramani, and Carl Rasmussen. The infinite hidden markov model. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 577–584. MIT Press, 2002.

[11] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *CoRR*, abs/1901.10912, 2019.

[12] V. Bentkus. A lyapunov-type bound in $\mathbb{R}^d$. *Theory of Probability and Its Applications (English translation)*, 49:311–323, 2005.

[13] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. An optimal high probability algorithm for the contextual bandit problem. *CoRR*, abs/1002.4058, 2010.

[14] D. Bouneffouf, I. Rish, and C. Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2020.

[15] Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, page 33–42, 1998.

[16] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 418–427. PMLR, 2019.

[17] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

[18] Xinshi Chen, Hanjun Dai, and Le Song. Particle flow Bayes' rule. In *Proceedings of the International Conference on Machine Learning*, pages 1022–1031, 2019.

[19] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *AISTATS 2011*, 2011.

[20] Hamid Dadkhahi and Sahand Negahban. Alternating linear bandits for online matrix-factorization recommendation. *CoRR*, abs/1810.09401, 2018.

[21] Varsha Dani, Thomas Hayes, and Sham Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pages 355–366, 2008.

[22] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Proceedings of the Machine Learning for Healthcare Conference*, pages 67–82, 2018.

[23] Jo Eidsvik, Tapan Mukerji, and Debarun Bhattacharjya. *Value of Information in the Earth Sciences: Integrating Spatial Modeling and Decision Analysis*. Cambridge University Press, 2015.

[24] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv e-prints*, page arXiv:0805.3415, May 2008.

[25] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

[26] Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. *Journal of Machine Learning Research*, pages 397–422, 2015.

[27] Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex bandit problems. In *Proceedings of the International Conference on Machine Learning*, pages 100–108, 2014.

[28] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition, 2020.

[29] Cédric Hartland, Nicolas Baskiotis, Sylvain Gelly, Michèle Sebag, and Olivier Teytaud. Change point detection and meta-bandits for online learning in dynamic environments. In *CAp 2007 : 9è Conférence francophone sur l'apprentissage automatique*, pages 237–250, July 2007.

[30] Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, and Amr Ahmed. Non-stationary off-policy optimization. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2494–2502. PMLR, 2021.

[31] Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. In *Advances in Neural Information Processing Systems*, pages 13423–13433, 2020.

[32] Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Mohammad Ghavamzadeh, and Craig Boutilier. Non-stationary latent bandits. *arXiv e-prints*, page arXiv:2012.00386, December 2020.

[33] Ronald Howard and James Matheson. Influence diagrams. In R. Howard and J. Matheson, editors, *The Principles and Applications of Decision Analysis*, volume II. Strategic Decisions Group, Menlo Park, CA, 1984.

[34] Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection, 2017.

[35] Jaya Kawale, Hung Bui, Branislav Kveton, Long Tran Thanh, and Sanjay Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems*, page 1297–1305, 2015.

[36] T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[37] Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

[38] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.

[39] Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems*, pages 2573–2583, 2018.

[40] Lihong Li. Generalized thompson sampling for contextual bandits. 2013.

[41] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. page 661–670, 2010.

[42] Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *CoRR*, abs/1812.10576, 2018.

[43] Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the Conference on Learning Theory*, pages 1739–1776, 2018.

[44] Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *Proceedings of the International Conference on Machine Learning*, pages 136–144, 2014.

[45] Gary Mavko, Tapan Mukerji, and Jack Dvorkin. *The Rock Physics Handbook: Tools for Seismic Analysis of Porous Media*. Cambridge University Press, 2nd edition, 2009.

[46] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, USA, 2000.

[47] Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A Bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

[48] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):949–1348, 2014.

[49] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis for thompson sampling with many actions. In *Advances in Neural Information Processing Systems*, 2018.

[50] Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, and Sanjay Shakkottai. Latent contextual bandits: A non-negative matrix factorization approach. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 518–527, 2017.

[51] Ross D. Shachter and David Heckerman. Thinking backward for knowledge acquisition. *AI Magazine*, 8(3):55–61, 1987.

[52] Nihal Sharma, Soumya Basu, Karthikeyan Shanmugam, and Sanjay Shakkottai. Warm starting bandits with side information from confounded data. *CoRR*, 2020.

[53] Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, page 974–980, 2015.

[54] Matthijs T. J. Spaan. *Partially Observable Markov Decision Processes*, pages 387–414. Springer Berlin Heidelberg, 2012.

[55] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[56] 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, 12 1933.

[57] Qing Wang, Chunqiu Zeng, Wubai Zhou, Tao Li, Larisa Shwartz, and Genady Ya. Grabarnik. Online interactive collaborative filtering using multi-armed bandit with dependent arms. *CoRR*, abs/1708.03058, 2017.

[58] Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal bandits with propagating inference. In *Proceedings of the International Conference on Machine Learning*, pages 5512–5520, 2018.

[59] Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the International Conference on Machine Learning*, page 1177–1184, 2009.

[60] Tong Yu, Branislav Kveton, Zheng Wen, Ruiyi Zhang, and Ole J. Mengshoel. Graphical models meet bandits: A variational Thompson sampling approach. In *Proceedings of the International Conference on Machine Learning*, pages 10902–10912, 2020.

[61] Se-Young Yun, Jun Hyun Nam, Sangwoo Mo, and Jinwoo Shin. Contextual multi-armed bandits under feature uncertainty. 2017.

[62] Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block mdps, 2020.

[63] Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3646–3653, 2016.

[64] Qian Zhou, XiaoFang Zhang, Jin Xu, and Bin Liang. Large-scale bandit approaches for recommender systems. In *Advances in Neural Information Processing Systems*, pages 811–821, 2017.

[65] Feiyun Zhu, Jun Guo, Ruoyu Li, and Junzhou Huang. Robust actor-critic contextual bandit for mobile health (MHealth) interventions. In *Proceedings of the ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, page 492–501, 2018.