Linearizing Contextual Bandits with Latent State Dynamics (Supplementary material)

Elliot Nelson[1] **Debarun Bhattacharjya**[1] **Tian Gao**[1] **Miao Liu**[1] **Djallel Bouneffouf**[1] **Pascal Poupart**[2]

[1]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[2]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada

A ONLINE EXPECTATION MAXIMIZATION FOR HIDDEN MARKOV MODELS

In sections (A.1)-(A.2) below, we describe the online EM algorithms used (by both L^{2}TS and L^{2}UCB) in our experiments.

These online EM algorithms involve updating the model posterior over the latent state with Bayes' rule,[1]

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

using the current parameter estimates $(\hat{\phi}^{(t-1)}, \hat{\theta}^{(t-1)})$. (These updates are shown in Eqs. (A.2) and (A.6) below, in the case of multinomial and Gaussian context distributions, respectively.)

In both cases, online EM uses a discount factor $\gamma_t \in (0,1)$ which is used to control the magnitude of parameter estimate updates over time. The rate at which γ_t approaches zero as $t \to \infty$ controls the discounting of previously observed context data. (In our experiments we use $\gamma_t = t^{-0.6}$.)

While we focus on Gaussian distributions in the case of continuous context data, the online EM algorithm of Cappé [2011] applies more generally to context distributions $p(x|z)$ in the exponential family.

A.1 MULTINOMIAL CONTEXT DISTRIBUTIONS

For multinomial context distributions with $x \in \{1, ..., X\}$, we define $\hat{\theta} = \{\hat{\nu}_{j,i}\}$ where $\hat{\nu}_{j,i} := p(x = i|z = j)$ satisfies $\sum_{i=1}^{X} \hat{\nu}_{j,i} = 1$. We use the algorithm of Mongillo and Deneve [2008] – reproduced in Eqs. (A.2)-(A.5) below – to implement the online EM update in L^{2}TS (Algorithm 1) and L^{2}UCB (Algorithm 2). We define OnlineEM$(x, \hat{\theta}^{(t-1)}, \hat{\phi}^{(t-1)}, \hat{p}_{t-1}, \hat{\psi}_{t-1})$ as the function which returns $(\hat{\theta}^{(t)}, \hat{\phi}^{(t)}, \hat{\psi}_t)$, where (in the categorical case) $\hat{\theta}^{(t)} = \{\hat{\nu}^{(t)}_{j,i}\}$, $\hat{\phi}^{(t)}$, and $\hat{\psi}_t = \{\hat{\rho}^{(t)}_{i,j,h}(k)\}$ are computed as in Eqs. (A.5), (A.4), and (A.3) respectively.

[1] The $\propto$ sign indicates equality up to a normalizing constant.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

$$\hat{p}_t(z) \propto \sum_{z'} \hat{p}_{t-1}(z') \hat{\phi}^{(t-1)}_{z,z'} \nu^{(t-1)}_{z,x_t} \tag{A.2}$$

$$\hat{\rho}^{(t)}_{i,j,h}(k) = \sum_{l} \Gamma_{l,h}(x_t) \Big((1-\gamma_t) \hat{\rho}^{(t-1)}_{i,j,l}(k) + \gamma_t \mathbf{1}(x_t = k) \mathbf{1}(i = l) \mathbf{1}(j = h) \hat{p}_{t-1}(l) \Big) \tag{A.3}$$

$$\text{where } \Gamma_{i,j}(x_t) = \frac{\hat{\phi}^{(t-1)}_{i,j} \hat{\nu}^{(t-1)}_{j,x_t}}{\sum_{i',j'} \hat{\phi}^{(t-1)}_{i',j'} \hat{\nu}^{(t-1)}_{j',x_t} \hat{p}_{t-1}(i')}$$

$$\hat{\phi}^{(t)}_{j,i} \propto \sum_{k=1}^{X} \sum_{h=1}^{Z} \hat{\rho}^{(t)}_{i,j,h}(k) \tag{A.4}$$

$$\hat{\nu}^{(t)}_{j,i} \propto \sum_{i,h=1}^{Z} \hat{\rho}^{(t)}_{i,j,h}(k) \tag{A.5}$$

In the updates to $\hat{p}_t$, $\hat{\phi}^{(t)}$, and $\hat{\nu}^{(t)}$ above, the $\propto$ sign indicates equality up to the normalizing factors required to ensure that $\sum_z \hat{p}_t(z) = 1$, $\sum_{z'} \hat{\phi}^{(t)}_{z',z} = 1$, or $\sum_{i=1}^{X} \hat{\nu}_{j,i} = 1$.

A.2 GAUSSIAN CONTEXT DISTRIBUTIONS

For Gaussian context distributions $p(x|z;\hat{\theta})$, the parameters are means and variances, $\hat{\theta} = \{\hat{\nu}_z, \hat{\Sigma}_z\}_1^Z$, conditional on each latent state z. In this case, we use Algorithm 1 of Cappé [2011] to implement the online EM parameter update in L^{2}TS. This algorithm is reproduced as follows, largely following the notation in Cappé [2011], with some modifications to maintain consistency with our notation in the main test.[2] We assume for simplicity that $x_t \in \mathbb{R}$ so that $\hat{\nu}^{(t)}_z$ is univariate. (The expressions in Cappé [2011] apply also to the multivariate case.)

We again define OnlineEM$(x, \hat{\theta}^{(t-1)}, \hat{\phi}^{(t-1)}, \hat{p}_{t-1}, \hat{\psi}_{t-1})$ as the function which returns $(\hat{\theta}^{(t)}, \hat{\phi}^{(t)}, \hat{\psi}_t)$, where now, in the Gaussian case, $\hat{\theta}^{(t)} = \{\hat{\nu}^{(t)}_z, \hat{\Sigma}^{(t)}_z\}$, $\hat{\phi}^{(t)}$, and $\hat{\psi}_t = \{\hat{\rho}^{(\phi)}_t(i,j,k), \hat{\boldsymbol{\rho}}^{(\theta)}_t(i,k)\}$ are computed as in Eqs. (A.12)-(A.13), (A.10), and (A.8)-(A.9), respectively. These updates involve the quadratic sufficient statistic, $\mathbf{s}(x) = [1, x, x^2]$, for context observations $x \sim p(\cdot|z;\theta^\star)$. In Eqs. (A.9) and (A.11) below, $\hat{\boldsymbol{\rho}}^{(\theta)}_t(i,k)$ shares the same vector dimension, which we indicate with bold symbols.

[2] In particular, Cappé [2011] uses $\hat{\phi}$ to denote the posterior probability vector which we call $\hat{p}$, and uses q to denote the latent transition probabilities $\hat{\phi}$.

$$\hat{p}_t(z) \propto \sum_{z'=1}^{Z} \hat{p}_{t-1}(z')\hat{\phi}^{(t-1)}_{z,z'} \frac{1}{\sqrt{2\pi\hat{\Sigma}^{(t-1)}_z}} \exp\left[-\left(x_t - \hat{\nu}^{(t-1)}_z\right)^2 \Big/ 2\hat{\Sigma}^{(t-1)}_z\right] \tag{A.6}$$

$$\hat{r}_t(z|z') = \frac{\hat{p}_{t-1}(z)\hat{\phi}^{(t-1)}_{z',z}}{\sum_{z''}\hat{p}_{t-1}(z'')\hat{\phi}^{(t-1)}_{z',z''}} \tag{A.7}$$

$$\hat{\rho}^{(\phi)}_t(i,j,k) = \gamma_t \mathbf{1}(j=k)\hat{r}_t(i|j) + (1-\gamma_t)\sum_{k'=1}^{Z}\hat{\rho}^{(\phi)}_{t-1}(i,j,k')\hat{r}_t(k'|k) \tag{A.8}$$

$$\hat{\boldsymbol{\rho}}^{(\theta)}_t(i,k) = \gamma_t\mathbf{1}(j=k)\mathbf{s}(x_t) + (1-\gamma_t)\sum_{k'=1}^{Z} \hat{\boldsymbol{\rho}}^{(\theta)}_{t-1}(i,k')\hat{r}_t(k'|k) \tag{A.9}$$

$$\hat{\phi}^{(t)}_{j,i} = \frac{\sum_{z=1}^{Z} \hat{\rho}^{(\phi)}_t(i,j,z)\hat{p}_t(z)}{\sum_{z',z=1}^{Z} \hat{\rho}^{(\phi)}_t(i,z',z)\hat{p}_t(z)} \tag{A.10}$$

$$\hat{\mathbf{S}}^{(\theta)}_t(i) = \sum_{k=1}^{Z} \hat{\boldsymbol{\rho}}^{(\theta)}_t(i,k)\hat{p}_t(k) \tag{A.11}$$

$$\hat{\nu}^{(t)}_z = \hat{S}^{(\theta)}_{t,1}(z)/\hat{S}^{(\theta)}_{t,0}(z) \tag{A.12}$$

$$\hat{\Sigma}^{(t)}_z = \hat{S}^{(\theta)}_{t,2}(z)/\hat{S}^{(\theta)}_{t,0}(z) - (\hat{\nu}^{(t)}_z)^2 \tag{A.13}$$

B EXPERIMENTS

In both L^{2}TS (Algorithm 1) and L^{2}UCB (Algorithm 2), and for all experiments, we use the following settings:

Online EM hyperparameters. We use $\gamma_t = t^{-0.6}$, following Cappé [2011].

Linear bandit hyperparameters. In L^{2}TS, we set $\tilde{\sigma}_r = 1$. In L^{2}UCB, we set $\alpha_{\text{UCB}} = 3$ for all experiments, which we found to improve convergence of regret compared to $\alpha_{\text{UCB}} = 1$. For both, we set $\lambda_\mu = 1$.

B.1 MULTINOMIAL CONTEXT DISTRIBUTIONS WITH BINARY REWARDS

Problem 1. Expressing the multinomial context distribution probabilities in matrix form, we set

$$p(x|z) = \begin{bmatrix} 0.05 & 0.05 & 0.45 & 0.45 \\ 0.45 & 0.45 & 0.05 & 0.05 \end{bmatrix}, \tag{B.1}$$

where $x \in \{1, ..., X\}$ with $X = 4$, and $z \in \{1, 2\}$.

Denoting the Bernoulli probabilities for (binary) reward values in matrix form, with actions $a \in \{1, 2\}$ and latent states $z \in \{1, 2\}$ indexing rows and columns, we set

$$p(r = 1|z, a) = \begin{bmatrix} 0.4 & 0.4 \\ 0.6 & 0.4 \end{bmatrix}. \tag{B.2}$$

The initial latent state was generated from a probability vector sampled from a (uniform) Dirichlet prior with concentration parameters $\alpha_z = 1$ for $z = 1, 2$.

Problem 2. In this problem, we set $(Z, X, K) = (4, 12, 8)$ with $p(x = i|z = j) = 1/3$ when $(i - j)\text{mod}X \in \{-1, 0, 1\}$, and zero otherwise. Reward probabilities $p(r = 1|z, a)$ are sampled uniformly in $(0, 1)$ for all (z, a). Latent states transition to the same state with probability 0.75, and to any other state with equal probabilities. For this task only, we omitted the optional reward update in L^{2}TS and L^{2}UCB, which we found to marginally increase regret. In this task, contexts x_t contain significantly more information about the current latent state z_t than do rewards r_t. The initial latent state was sampled from a uniform probability distribution.

Online EM initialization. The sufficient statistics introduced in Appendix A.1 were initialized at $\hat{\rho}^{(0)}_{i,j,h}(k) = 0.01$ for all (i, j, h) and all $k \in \{1, ..., X\}$. The initial latent state probability vector $\hat{p}_0(z)$ was sampled randomly from the uniform distribution over probability vectors.

B.2 MINING APPLICATION DETAILS

We assume a Gaussian reward model, $p(r|z, a) = \mathcal{N}(\hat{\mu}^{(a)}_z, \tilde{\sigma}^2_r)$. The variance $\tilde{\sigma}^2_r$ is a fixed hyperparameter, which we equate with the hyperparameter $\tilde{\sigma}_r$ in Algorithm 1 used for Thompson sampling. (This hyperparameter is the variance of the implicit Gaussian reward likelihood used in L^{2}TS to update the multivariate Gaussian posterior over $\mu^{(a)}$.)

Online EM initialization. The sufficient statistics introduced in Appendix A.2 were initialized at $\hat{\rho}^{(\phi)}_0(i, j, k) = 1$, $\hat{\boldsymbol{\rho}}^{(\theta)}_0(i, k) = [1, 1, 1]$ for all (i, j, k). The initial latent state probability vector $\hat{p}_0(z)$ was set to a uniform distribution.

Numerical Details of Application. We model this application using three latent rock classes $z = 1, 2, 3$, two mining actions $a = 1, 2$, and Gaussian contextual observations for hand-held x-ray flourescent meter (XMET) measurements.

To model $p(x|z)$, we use the approach and numbers in Eidsvik et al. [2015] for how the continuous-valued XMET observations depend on the latent rock class. The ore grade o and the observed continuous XMET observation x follow Gaussian distributions as follows:

$$o = \beta_0 + \beta_1 z + N(0, \sigma^2);\ x = o + N(0, \tau^2), \tag{B.3}$$

where $\beta_0 = -0.18$ and $\beta_1 = 1.32$ are regressions coefficients. The latter coefficient signifies that a higher rock class results in higher ore grade. $\sigma = 0.62$ captures the uncertainty in the ore grade and $\tau = 0.45$ captures the quality of the observed XMET. These numbers are directly from Eidsvik et al. [2015].

For the reward distribution $p(r|a, z)$, we assume the profit depends on a revenue factor (r_f) per ore grade from mined ore as well as both fixed (c_f) and uncertain (variable) costs (c_v):

$$\text{Profit}(a, z) = o(z) * r_f(a) - c_f(a) - c_v(a), \tag{B.4}$$

where $c_v(a) \sim N(0, \sigma^2_c)$. We choose numbers such that action $a = 1$ has a higher revenue factor and more fixed cost but less variable cost compared to action $a = 2$. Note that the profit is Gaussian as it is linear in Gaussian random variables.

We choose a transition matrix over latent states that favors the diagonal, because spatial modeling in general heavily uses covariance related concepts (such as variograms) where regions that are geographically closer are more correlated. In our first experiment, we choose the following matrix, where rows from top to bottom are for $z = 1, 2, 3$:

$$p(z_t|z_{t-1}) = \begin{bmatrix} 0.7 & 0.25 & 0.05 \\ 0.25 & 0.5 & 0.25 \\ 0.05 & 0.25 & 0.7 \end{bmatrix} \tag{B.5}$$

In our second experiment, we choose a matrix for which latent state changes are very rare, occuring only every $O(50)$ steps:

$$p(z_t|z_{t-1}) = \begin{bmatrix} 0.98 & 0.01 & 0.01 \\ 0.01 & 0.98 & 0.01 \\ 0.01 & 0.01 & 0.98 \end{bmatrix} \tag{B.6}$$

In both cases, the initial latent state was generated from a probability vector sampled from a Dirichlet prior with concentration parameters $\alpha_z = 1$ for $z = 1, 2, 3$ (i.e. a uniform prior over probability vectors).

B.3 PARAMETER ESTIMATION ERROR

The gap between L^{2}TS and the corresponding oracle variant in Figure 3 (which conditions on the true parameters $\theta^\star$, $\phi^\star$) is a consequence of parameter estimation error. In Figure B.1 below we show parameter estimation error of online EM, when used by L^{2}TS for the mining application described above. We show mean squared errors averaged over 10 episodes with different randomly generated ground truth transition matrices (with each column sampled from a uniform distribution over probability vectors), as well as different randomly generated mean values $\mathbb{E}[x|z] \sim \mathcal{N}(0, 1)$ for the Gaussian conditional distributions $p(x|z)$. (Otherwise, we use the same environment parameters as described above.)

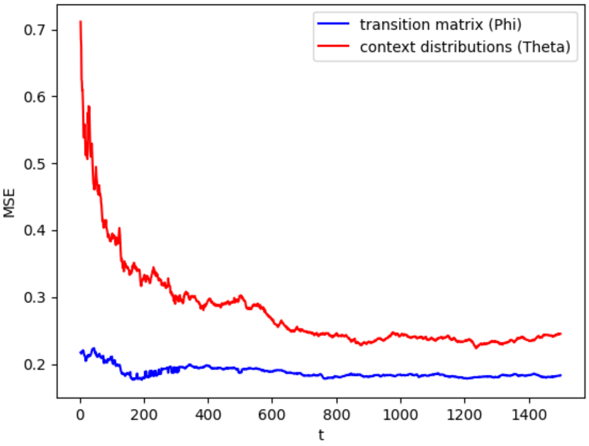

Figure B.1: Mean squared error (MSE) of model estimates for the latent transition matrix ($||\hat{\phi}^{(t)} - \phi^\star||_2^2$) and context distributions ($||\hat{\theta}^{(t)} - \theta^\star||_2^2$) used by L^{2}TS on the mining application, averaged over 10 different true transition matrices and context distributions.

B.4 BASELINE DETAILS.

We allowed the discounted Thompson Sampling (dTS) algorithm to access the true transition matrix to set its discount factor to $\gamma = Z^{-1} \sum_z \phi^\star_{z,z}$.

For umTS [Hong et al., 2020], we used $N = 100$ particles with a minimum effective sample size $ESS_{min} = 20$ for particle resampling.

For Exp4.P [Beygelzimer et al., 2010], we pretrained 10 expert modules with linear regression (in the case of categorical contexts and rewards) or MLP classification[3] (in the case of Gaussian context and rewards) to classify observations x into corresponding optimal actions, based on 1000 samples of contexts $x \sim p(x) = \sum_z p(x|z; \theta^\star)p_i(z)$ and action-wise rewards $r_a \sim p(r|a) = \sum_z p(r|z, a)p_i(z)$. For each expert i, we used a different categorical distribution $p_i(z)$ obtained by sampling a uniform distribution over the Z-simplex. (We found comparable performance when increasing to 50 experts.) The Exp4.P algorithm then learns to give greater weight to experts who were trained on distributions $p_i(z)$ which assign higher probability to recently occurring latent states in the non-stationary environment.

C 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. (4), and (as discussed in Section 3.3), used by Algorithm 1 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, and thus replace the model posterior $\hat{p}_t$ with the "true" posterior $p^\star_t$ as defined in Eq. (2). 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. (4)) and true mean reward parameters as

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

where

$$g^{(a)} := f^{(a)}_\mu - B^{(a)}\mu^{(a)}_\star = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p^\star_t \left(r_t - (p^\star_t)^\top \mu^{(a)}_\star \right). \tag{C.2}$$

[3] We use the default architecture settings specified at https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.

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

$$g_z^{(a)} = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p^\star(z_t = z|x_{1:t}) \Big(r_t - \sum_{z'} p^\star(z_t = z'|x_{1:t}) (\mu_\star^{(a)})_{z'} \Big). \tag{C.3}$$

$$B_{zz'}^{(a)} = \sum_{t=1}^{T} \mathbf{1}(a_t = a) p^\star(z_t = z|x_{1:t}) p^\star(z_t = z'|x_{1:t}). \tag{C.4}$$

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

$$p_{t:t'}(z) := p(z_{t'} = z|x_{t:t'}) \tag{C.5}$$

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 C.1 we derive several intermediate results using a contraction property [Boyen and Koller, 1998] 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 C.2 and Appendix C.3, we use the results of Appendix C.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 C.4 we convert the element-wise bound on $B^{(a)}$ into a bound on the largest eigenvalue of $(B^{(a)})^{-1}$.
- In Appendix C.5 we combine the results of the previous two sections to obtain the final high-probability bound on the estimator error $\hat{\mu}^{(a)} - \mu_\star^{(a)}$.

C.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* [Boyen and Koller, 1998] of a transition matrix,

$$\gamma_\phi := \min_{z_1, z_2} \sum_z \min(\phi_{z,z_1}, \phi_{z,z_2}). \tag{C.6}$$

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 Boyen and Koller [1998] to prove a contraction theorem for the KL divergence between two different distributions:

Theorem C.1 (Theorem 3 in Boyen and Koller [1998]). *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*

$$D_{KL}[p||q] \leq (1 - \gamma_\phi) D_{KL}[p_0||q_0], \tag{C.7}$$

with the minimal mixing rate γ_ϕ defined in Eq. (C.6).

We will also make use of the fact [Boyen and Koller, 1998] that conditioning on additional data reduces the KL divergence between different distributions, in expectation:

Lemma C.2. *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 $\ell(x|z)$,*

$$p(z|x) = p(z)\ell(x|z)/p(x), \;\; q(z|x) = q(z)\ell(x|z)/q(x), \tag{C.8}$$

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

$$\mathbb{E}_{x\sim p(x)}[D_{KL}[p(z|x)||q(z|x)]] \leq D_{KL}[p(z)||q(z)]. \tag{C.9}$$

Proof. Using Eq. (C.8), we have

$$\begin{aligned}
\mathbb{E}_{x\sim p(x)}[D_{KL}[p(z|x)||q(z|x)]] &= \mathbb{E}_{x\sim p(x)}\left[\sum_z \frac{p(z)\ell(x|z)}{p(x)} \log\left(\frac{p(z)\ell(x|z)}{p(x)}\frac{q(x)}{q(z)\ell(x|z)}\right)\right] \\
&= \mathbb{E}_{x\sim p(x)}\left[\sum_z \frac{p(z)\ell(x|z)}{p(x)} \left(\log\frac{p(z)}{q(z)} - \log\frac{p(x)}{q(x)}\right)\right] \\
&= \sum_z p(z)\log\frac{p(z)}{q(z)}\mathbb{E}_{x\sim p(x)}[\ell(x|z)/p(x)] - \mathbb{E}_{x\sim p(x)}\left[\frac{\sum_z p(z)\ell(x|z)}{p(x)}\log\frac{p(x)}{q(x)}\right] \\
&= D_{KL}[p(z)||q(z)] - D_{KL}[p(x)||q(x)].
\end{aligned} \tag{C.10}$$

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. (C.9). □

Eq. (C.7) and Eq. (C.9) 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 C.3. *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

$$\mathbb{E}_{x\sim p(\cdot|z';\theta), z'\sim\Phi\tilde{q}_0}[D_{KL}[\tilde{q}||q]] \leq (1-\gamma_\phi) D_{KL}[\tilde{q}_0||q_0], \tag{C.11}$$

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. (C.9) 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. (C.7), we recover Eq. (C.11). □

Note that Eq. (C.11) – and consequently also Eqs. (C.14) and (C.17) 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. (C.11) can be applied recursively to show that the KL divergence contracts exponentially as the two distributions are propagated forward in time:

Lemma C.4. *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*

$$q_t(z') := \sum_z q_0(z) p(z_t = z'|x_{1:t}, z_0 = z), \tag{C.12}$$

$$\tilde{q}_t(z') := \sum_z \tilde{q}_0(z) p(z_t = z'|x_{1:t}, z_0 = z), \tag{C.13}$$

satisfy

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}[D_{KL}[\tilde{q}_t||q_t]] \leq e^{-\gamma_\phi t} D_{KL}[\tilde{q}_0||q_0], \tag{C.14}$$

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. (C.11) to the transition at time t, with priors $(\tilde{q}_0, q_0) \to (\tilde{q}_{t-1}, q_{t-1})$ in Eq. (C.11) determined by a fixed sequence $x_{1:t-1}$ of preceding data, we have

$$\mathbb{E}_{x_t|x_{1:t-1},z_0\sim\tilde{q}_0}[D_{KL}[\tilde{q}_t||q_t]] \leq (1-\gamma_\phi) D_{KL}[\tilde{q}_{t-1}||q_{t-1}], \tag{C.15}$$

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

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}[D_{KL}[\tilde{q}_t||q_t]] \leq (1-\gamma_\phi)^t D_{KL}[\tilde{q}_0||q_0], \tag{C.16}$$

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. (C.14). □

Note that Eq. (C.14) is a conservative bound, for two reasons: (1) If there exist pairs of states (z_1, z_2) in Eq. (C.6) – 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. (C.9) is a weaker bound than Eq. (C.10), 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 C.2 – are very different, making observations x highly informative about z.

Eq. (C.14) can be converted into a bound on the expected total variation distance, or 1-norm, between two posteriors:

Corollary C.4.1. *The* 1*-norm difference between two distributions $(\tilde{q}_t, q_t)$ over the state z_t, as defined in Eqs.* (C.12)-(C.13)*, satisfies the upper bound*

$$\mathbb{E}_{x_{1:t}|z_0\sim\tilde{q}_0}\left[\sum_z |\tilde{q}_t(z) - q_t(z)|\right] \leq e^{-\frac{1}{2}\gamma_\phi t}\sqrt{2D_{KL}[\tilde{q}_0||q_0]}. \tag{C.17}$$

Proof. Pinsker's inequality states that for any two probability distributions $\tilde{q}$ and q, the 1-norm and KL divergence satisfy $||\tilde{q} - q||_1 \leq \sqrt{2D_{KL}[\tilde{q}||q]}$.[4] 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. (C.14), we arrive at Eq. (C.17). □

C.2 PARTIAL BOUND ON THE ESTIMATOR ERROR

In this section, we compute the variance of the vector $g^{(a)}$ defined in Eq. (C.2), across different reward and observation histories. We show that the variance converges to zero 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 $\hat{\mu}^{(a)} - \mu_\star^{(a)} = (B^{(a)})^{-1} g^{(a)}$.

[4] 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 $D_{KL}[q||\tilde{q}]$.

Lemma C.5. *When the ground truth parameters (θ, Φ) are known, each element of $g^{(a)}$, Eq.* (C.2)*, satisfies the upper bound*

$$(g_z^{(a)}/T)^2 \leq \frac{1}{\delta \cdot T}\Big(\sigma_{\text{eq}}^2 + ||\mu_\star^{(a)}||_1^2 \frac{4}{\gamma_\phi}\big(1 + \log \zeta_\phi\big)\Big) \tag{C.18}$$

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

$$\sigma_{\text{eq}}^2 := \max_a \sum_z \rho_{\text{eq}}^{(\phi)}(z) \text{Var}[r|z, a] \tag{C.19}$$

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

$$\tau^\star := \min_{\tau \in \mathbb{N}} |\log D_\phi(\tau) - \gamma_\phi \tau|, \tag{C.20}$$

where

$$D_\phi(\tau) := \max_z \max_{t \geq 1} \mathbb{E}_{x_{1:t+\tau}}\left[D_{KL}[p(z_{t+\tau}|x_{1:t+\tau})||p(z_{t+\tau}|z_t = z; x_{1:t+\tau})]\right] \tag{C.21}$$

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:

$$\begin{aligned}
\mathbb{E}[g_z^{(a)}|a_{1:T}] &= \mathbb{E}_{x_{1:T}}[g_z^{(a)}|x_{1:T}, a_{1:T}] \\
&= \mathbb{E}_{x_{1:T}}\Big[\sum_{t=1}^T \delta_t^{(a)} p(z_t = z|x_{1:t}) \mathbb{E}[r_t|x_{1:T}, a_t = a] \\
&\qquad - \sum_{t=1}^T \delta_t^{(a)} p(z_t = z|x_{1:t}) \sum_{z'} p(z_t = z'|x_{1:t})(\mu_\star^{(a)}) r_{z'}\Big] \\
&= \sum_{t=1}^T \delta_t^{(a)} \mathbb{E}_{x_{1:T}}\left[p(z_t = z|x_{1:t}) \sum_{z'} (p(z_t = z'|x_{1:T}) - p(z_t = z'|x_{1:t}))(\mu_\star^{(a)})_{z'}\right] \\
&= \sum_{t=1}^T \delta_t^{(a)} \sum_{z'} (\mu_\star^{(a)})_{z'} \cdot \mathbb{E}_{x_{1:t}}\left[p(z_t = z|x_{1:t})(\mathbb{E}_{x_{t+1:T}}[p(z_t = z'|x_{1:T})] - p(z_t = z'|x_{1:t}))\right] \\
&= \sum_{t=1}^T \delta_t^{(a)} \sum_{z'} (\mu_\star^{(a)})_{z'} \cdot \mathbb{E}_{x_{1:t}}[p(z_t = z|x_{1:t})(p(z_t = z|x_{1:t}) - p(z_t = z|x_{1:t}))] = 0.
\end{aligned} \tag{C.22}$$

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

$$\eta_t^{(a)} := r_t - \sum_{z'} p(z_t = z'|x_{1:t})(\mu_\star^{(a)})_{z'} = r_t - p_t^\top \mu_\star^{(a)}, \tag{C.23}$$

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))

$$\mathbb{E}[g_{z_1}^{(a)} g_{z_2}^{(a)}|x_{1:T}, a_{1:T}] = \sum_{t,t'} \delta_t^{(a)} \delta_{t'}^{(a)} p(z_t = z_1|x_{1:t}) p(z_{t'} = z_2|x_{1:t'}) \cdot \mathbb{E}[\eta_t^{(a)} \eta_{t'}^{(a)}|x_{1:T}, a_t = a_{t'} = a]. \tag{C.24}$$

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

$$\begin{aligned}
&(\text{for } t \neq t') \\
&\mathbb{E}[\eta_t\eta_{t'}|x_{1:T}, a_t = a_{t'} = a] = \\
&\quad \sum_{z,z'} (\mu_\star^{(a)})_z(\mu_\star^{(a)})_{z'} \big[p(z_t = z, z_{t'} = z'|x_{1:T}) - p(z_t = z|x_{1:t})p(z_{t'} = z'|x_{1:T}) \\
&\qquad - p(z_t = z|x_{1:T})p(z_{t'} = z'|x_{1:t'}) + p(z_t = z|x_{1:t})p(z_{t'} = z'|x_{1:t'})\big]. \qquad \text{(C.25)}
\end{aligned}$$

When $t = t'$ we have

$$\begin{aligned}
\mathbb{E}[\eta_t^2|x_{1:T}, a_t = a] &= \sum_z p(z_t = z|x_{1:T})((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) \\
&\quad - 2\Big(\sum_z p(z_t = z|x_{1:t})(\mu_\star^{(a)})_z\Big)\Big(\sum_{z'} p(z_t = z'|x_{1:T})(\mu_\star^{(a)})_{z'}\Big) \\
&\quad + \Big(\sum_z p(z_t = z|x_{1:t})(\mu_\star^{(a)})_z\Big)^2, \qquad \text{(C.26)}
\end{aligned}$$

where

$$\sigma_z^{(a)} := \mathbb{E}_{r\sim p(\cdot|z,a)}[r^2] - \mathbb{E}_{r\sim p(\cdot|z,a)}[r]^2 = \mathbb{E}_{r\sim p(\cdot|z,a)}[r^2] - [(\mu_\star^{(a)})_z]^2. \qquad \text{(C.27)}$$

We now take the expectation over $x_{1:T}$. Because Eq. (C.24) 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:

$$\begin{aligned}
&\mathbb{E}[g_{z_1}^{(a)}g_{z_2}^{(a)}|a_{1:T}] = \mathbb{E}_{x_{1:T}}\Big[\mathbb{E}[g_{z_1}^{(a)}g_{z_2}^{(a)}|x_{1:T}, a_{1:T}]\Big] \\
&\quad = 2\sum_{t,t'>t} \delta_t^{(a)}\delta_{t'}^{(a)}\mathbb{E}_{x_{1:t'}}\Big[p(z_t = z_1|x_{1:t})p(z_{t'} = z_2|x_{1:t'}) \cdot \mathbb{E}_{x_{t'+1:T}}[\mathbb{E}[\eta_t\eta_{t'}|x_{1:T}, a_t = a_{t'} = a]]\Big] \\
&\qquad + \sum_t \delta_t^{(a)}\mathbb{E}_{x_{1:t}}\big[p(z_t = z_1|x_{1:t})p(z_t = z_2|x_{1:t}) \cdot \mathbb{E}_{x_{t+1:T}}[\mathbb{E}[\eta_t^2|x_{1:T}, a_t = a]]\big] \qquad \text{(C.28)}
\end{aligned}$$

where we have decomposed the double sum over time as $\sum_{t,t'} = \sum_{t=t'} + 2\sum_{t,t'>t}$. Using Eq. (C.25) for the $t < t'$ terms, we have

$$\begin{aligned}
&(\text{for } t < t')\ \mathbb{E}_{x_{t'+1:T}}[\mathbb{E}[\eta_t\eta_{t'}|x_{1:T}, a_t = a_{t'} = a]] \\
&\quad = \sum_{z,z'} (\mu_\star^{(a)})_z(\mu_\star^{(a)})_{z'} \big[\mathbb{E}_{x_{t'+1:T}}[p(z_t = z, z_{t'} = z'|x_{1:T})] \\
&\qquad - p(z_t = z|x_{1:t})\mathbb{E}_{x_{t'+1:T}}[p(z_{t'} = z'|x_{1:T})] \\
&\qquad - \mathbb{E}_{x_{t'+1:T}}[p(z_t = z|x_{1:T})]p(z_{t'} = z'|x_{1:t'}) \\
&\qquad + p(z_t = z|x_{1:t})p(z_{t'} = z'|x_{1:t'})\big] \\
&\quad = \sum_{z,z'} (\mu_\star^{(a)})_z(\mu_\star^{(a)})_{z'} \big[p(z_t = z, z_{t'} = z'|x_{1:t'}) - p(z_t = z|x_{1:t})p(z_{t'} = z'|x_{1:t'}) \\
&\qquad - p(z_t = z|x_{1:t'})p(z_{t'} = z'|x_{1:t'}) + p(z_t = z|x_{1:t})p(z_{t'} = z'|x_{1:t'})\big] \\
&\quad = \sum_{z,z'} (\mu_\star^{(a)})_z(\mu_\star^{(a)})_{z'} p(z_t = z|x_{1:t'})\big(p(z_{t'} = z'|z_t = z, x_{1:t'}) - p(z_{t'} = z'|x_{1:t'})\big) \qquad \text{(C.29)}
\end{aligned}$$

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. (C.26) for the $t' = t$ terms, we have

$$\begin{aligned}
\mathbb{E}[\eta_t^2|x_{1:t}, a_t = a] &= \mathbb{E}_{x_{t+1:T}}[\mathbb{E}[\eta_t^2|x_{1:T}, a_t = a]] \\
&= \sum_z p(z_t = z|x_{1:t})((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2) - \Big(\sum_z p(z_t = z|x_{1:t})(\mu_\star^{(a)})_z\Big)^2 \\
&\leq \sum_z p(z_t = z|x_{1:t})((\sigma_z^{(a)})^2 + [(\mu_\star^{(a)})_z]^2). \qquad \text{(C.30)}
\end{aligned}$$

Substituting Eqs. (C.29) and (C.30) into Eq. (C.28), taking the absolute value to obtain an upper bound, using $p(z_t = z_1|x_{1:t})p(z_{t'} = z_2|x_{1:t'}) \leq 1$ and $p(z_t = z|x_{1:t'}) \leq 1$ to simplify the expression, using the fact that $\mathbb{E}_{x_{1:t}}[p(z_t = z|x_{1:t})] = \rho^{(\phi)}_{\text{eq}}(z)$ in the $t = t'$ contribution, and setting $z_1 = z_2$ for simplicity, we have

$$\begin{aligned}\text{Var}[g^{(a)}_{z_1}] &\leq T \sum_z \rho^{(\phi)}_{\text{eq}}(z)((\sigma^{(a)}_z)^2 + [(\mu^{(a)}_\star)_z]^2) \\ &\quad + \sum_{z,z'} |(\mu^{(a)}_\star)_z (\mu^{(a)}_\star)_{z'}| \times 2 \sum_{t,t'>t} \mathbb{E}_{x_{1:t'}} \left[|p(z_{t'} = z'|z_t = z, x_{1:t'}) - p(z_{t'} = z'|x_{1:t'})|\right]\end{aligned} \tag{C.31}$$

We have also used $\delta^{(a)}_t \delta^{(a)}_{t'} \leq 1$ and have removed the action-conditioning on $\text{Var}[g^{(a)}]$, since after setting $\delta^{(a)}_t \delta^{(a)}_{t'} \leq 1$ 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 C.4.1 to bound the expectation value over $x_{t+\tau_1+1:t'}$:

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

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. (C.21). 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:

$$\begin{aligned}\text{Var}[g^{(a)}_{z_1}] &\leq T \sum_z \rho^{(\phi)}_{\text{eq}}(z)((\sigma^{(a)}_z)^2 + [(\mu^{(a)}_\star)_z]^2) \\ &\quad + 2||\mu^{(a)}_\star||^2_1 \sum_{t,t'>t} \left[\mathbf{1}(t' - t \leq \tau_0) + \mathbf{1}(t' - t > \tau_0) e^{-\frac{1}{2}\gamma_\phi(t'-(t+\tau_1))} \sqrt{2D_\phi(\tau_1)}\right].\end{aligned} \tag{C.32}$$

Here, we have used the fact that

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

and (in the $t' - t \leq \tau_0$ term) the fact that the difference of probabilities in Eq. (C.31) 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. (C.32), we have

$$\text{Var}[g^{(a)}_{z_1}] \leq T \sum_z \rho^{(\phi)}_{\text{eq}}(z)((\sigma^{(a)}_z)^2 + [(\mu^{(a)}_\star)_z]^2) + 2||\mu^{(a)}_\star||^2_1 \left(T\tau_0 + \frac{2T}{\gamma_\phi} e^{-\frac{1}{2}\gamma_\phi(\tau_0-\tau_1)} \sqrt{2D_\phi(\tau_1)}\right). \tag{C.33}$$

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 $\gamma_\phi \tau_1^\star = \log D_\phi(\tau_1^\star) := \log \zeta_\phi$. 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:

$$\mathrm{Var}[g_z^{(a)}] \leq T\Big(\sigma_{\mathrm{eq}}^2 + ||\mu_\star^{(a)}||_1^2 \frac{4}{\gamma_\phi}\big(1 + \log \zeta_\phi\big)\Big), \tag{C.34}$$

where

$$\sigma_{\mathrm{eq}}^2 := \max_a \sum_z \rho_{\mathrm{eq}}^{(\phi)}(z)(\sigma_z^{(a)})^2. \tag{C.35}$$

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

$$|g_z^{(a)} - \mathbb{E}[g_z^{(a)}]| < \sqrt{\frac{\mathrm{Var}[g_z^{(a)}]}{\delta}} \tag{C.36}$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$. Recalling from Eq. (C.22) that $\mathbb{E}[g_z^{(a)}] = 0$, we recover Eq. (C.18) above. □

C.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 $\mathbb{E}[B_{z,z'}^{(a)}]$ 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[5]) an optimal action

$$a_t^\star := \mathrm{argmax}_a \sum_z p^\star(z_t = z | x_{1:t}) \mu_\star^{(a)}, \tag{C.37}$$

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

Lemma C.6. *Assuming that at any t the optimal action given $x_{1:t}$, Eq.* (C.37)*, 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*

$$\frac{1}{T}\mathbb{E}[B^{(a)}(T)] \succcurlyeq \pi_{\min} \bar{B}^{(a)}(T), \tag{C.38}$$

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

$$\bar{B}_{zz'}^{(a)}(T) := \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{x_{1:t}}\left[\mathbf{1}(a = a_t^\star) p(z_t = z | x_{1:t}) p(z_t = z' | x_{1:t})\right]. \tag{C.39}$$

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

$$\begin{aligned} \mathbb{E}[B_{zz'}^{(a)}(T)] &= \sum_{t=1}^T \mathbb{E}_{x_{1:t}}\left[\mathbb{E}_{r_{1:t-1}, a_{1:t-1} | x_{1:t}}[\mathbf{1}(a_t = a)] p(z_t = z | x_{1:t}) p(z_t = z' | x_{1:t})\right] \\ &= \sum_{t=1}^T \mathbb{E}_{x_{1:t}}\left[P_\pi(a_t = a | x_{1:t}) p(z_t = z | x_{1:t}) p(z_t = z' | x_{1:t})\right]. \end{aligned} \tag{C.40}$$

[5] We restore the * notation in Eq. (C.37) to denote this.

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

$$P_\pi(a_t = a|x_{1:t}) := \mathbb{E}_{r_{1:t-1}, a_{1:t-1}|x_{1:t}}[\mathbf{1}(a_t = a)] \tag{C.41}$$

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,

$$P_\pi(a_t = a|x_{1:t}) \geq \pi_{\min} \cdot \mathbf{1}(a = a_t^\star), \tag{C.42}$$

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[6] for any $x_{1:t}$, Eq. (C.42) 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 $\mathbb{E}[B^{(a)}_{zz'}(T)]$ in Eq. (C.40) we see that

$$\mathbb{E}[B^{(a)}(T)] \succcurlyeq \pi_{\min} \cdot T \cdot \bar{B}^{(a)}(T), \tag{C.43}$$

with $\bar{B}^{(a)}(T)$ defined in Eq. (C.39). Hence we recover the matrix lower bound Eq. (C.38) above. □

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

Lemma C.7. *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*

$$\mathrm{Var}\left[\frac{1}{T} B^{(a)}_{zz'}(T)\right] \leq \frac{2}{\gamma_\phi T}(\kappa + \log\log(1/\rho_{\min})), \tag{C.44}$$

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

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

$$\begin{aligned}\mathrm{Var}[B^{(a)}_{zz'}|a_{1:T}] &= \mathbb{E}_{x_{1:T}}[(B^{(a)}_{zz'})^2|a_{1:T}] - \mathbb{E}_{x_{1:T}}[B^{(a)}_{zz'}|a_{1:T}]^2 \\ &= \sum_{t,t'} \delta_t^{(a)} \delta_{t'}^{(a)} \Big(\mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')p_{1:t'}(z)p_{1:t'}(z')] - \mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')]\mathbb{E}_{x_{1:t'}}[p_{1:t'}(z)p_{1:t'}(z')]\Big).\end{aligned} \tag{C.45}$$

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

$$\delta p_t(z;\tau) := p_{1:t}(z) - p_{t-\tau+1:t}(z), \tag{C.46}$$

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

$$p_{1:t'}(z)p_{1:t'}(z') = (p_{t+1:t'}(z') + \delta p_{t'}(z;t'-t))(p_{t+1:t'}(z') + \delta p_{t'}(z';t'-t)), \tag{C.47}$$

Using Corollary C.4.1 to bound the expectation over $x_{t+1:t'}$, and using the fact that

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

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

we have

$$\mathbb{E}_{x_{1:t'}}\Big[\sum_{z}|\delta p_{t'}(z;\tau)|\Big] \leq e^{-\frac{1}{2}\gamma_\phi\tau}\mathbb{E}_{x_{1:t}}[\sqrt{2D_{KL}[p_{1:t}||\rho_{\rm eq}^{(\phi)}}] \leq e^{-\frac{1}{2}\gamma_\phi\tau}\sqrt{2\log(1/\rho_{\min})} := u(\tau). \tag{C.48}$$

Thus, for $t' > t$,

$$\begin{aligned}|\mathbb{E}_{x_{1:t'}}[p_{1:t'}(z)p_{1:t'}(z')] - \mathbb{E}_{x_{t+1:t'}}[p_{t+1:t'}(z)p_{t+1:t'}(z')]| &\leq \mathbb{E}_{x_{1:t'}}[|\delta p_{t'}(z;t'-t)|] + \mathbb{E}_{x_{1:t'}}[|\delta p_{t'}(z';t'-t)|] \\ &\quad + \mathbb{E}_{x_{1:t'}}[|\delta p_{t'}(z;t'-t)|\cdot|\delta p_{t'}(z';t'-t)|] \\ &\leq 3u(t'-t),\end{aligned} \tag{C.49}$$

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. (C.47) again for the first term in Eq. (C.45), we have

$$\mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')p_{1:t'}(z)p_{1:t'}(z')] \leq \mathbb{E}_{x_{1:t}}[p_{1:t}(z)p_{1:t}(z')]\cdot\mathbb{E}_{t+1:t'}[p_{t+1:t'}(z)p_{t+1:t'}(z')] + 3u(t'-t). \tag{C.50}$$

Here we have used the fact that $p_{1:t}(z)p_{1:t}(z') \leq 1$ to simplify the last term. Combining Eq. (C.49) and (C.50), we have (for $t' > t$)

$$|\mathbb{E}_{x_{1:t'}}[p_{1:t}(z)p_{1:t}(z')p_{1:t'}(z)p_{1:t'}(z')] - \mathbb{E}_{x_{1:t}}[p_{1:t}(z)p_{1:t}(z')]\mathbb{E}_{x_{1:t'}}[p_{1:t'}(z)p_{1:t'}(z')]| \leq 6u(t'-t). \tag{C.51}$$

As in Lemma C.5, we now introduce a free parameter τ_0, and break the sum in Eq. (C.45) into a contributions from small $|t'-t|$ (where the difference in Eq. (C.45) may be large but cannot exceed one) and large $|t'-t|$ (where the upper bound on the difference in Eq. (C.45) is strong). The variance $\mathrm{Var}[B_{z,z'}]$, Eq. (C.45), can then be upper bounded:

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

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

$$\mathrm{Var}[B_{zz'}^{(a)}] \leq \sum_{t,t'}\mathbf{1}(|t'-t|\leq\tau_0) + 2\sum_{t,t'}\mathbf{1}(t'-t>\tau_0)6u(t'-t). \tag{C.52}$$

Here, we have also used the symmetry of Eq. (C.45) under exchange of t and t' to sum only over $t' > t$. The bound on $\mathrm{Var}[B_{z,z'}]$ becomes

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

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

$$\tau_0^\star := \frac{1}{\gamma_\phi}\log\left(72\log(1/\rho_{\min})\right), \tag{C.53}$$

for which the upper bound becomes

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

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

Note that the unusual log-log dependence in Eq. (C.44) originates in the exponential contraction in Eq. C.14, 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 C.8. *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*

$$\frac{1}{T}|B^{(a)}_{zz'}(T) - \mathbb{E}[B^{(a)}_{zz'}(T)]| \leq \sqrt{\frac{1}{\delta}\frac{2}{\gamma_\phi T}(\kappa + \log\log(1/\rho_{\min}))} \tag{C.54}$$

where $\kappa \approx 6.78$, 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. (C.44) to upper bound the variance, we recover Eq. (C.54) above. □

C.4 BOUND ON COVARIANCE MATRIX EIGENVALUES

In Appendix C.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. (C.1). In this section, we show that an element-wise bound such as Eq. (C.54) can be converted to an eigenvalue bound which can be applied to the inverse matrix.

Lemma C.9. *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*

$$\lambda_1 \geq \bar{\lambda}_1 - ZU_\delta \tag{C.55}$$

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

$$\lambda_1 \geq \bar{\lambda}_1 + \lambda_1^{(\Delta)}. \tag{C.56}$$

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

$$|\lambda_i| \leq \sum_j |A_{ij}| \tag{C.57}$$

Applying Eq. (C.57) to any eigenvalue $\lambda_z^{(\Delta)}$ of ΔM, we have

$$|\lambda_z^{(\Delta)}| \leq \sum_{z'} |\Delta M_{zz'}| \leq ZU_\delta. \tag{C.58}$$

Since Eq. (C.58) 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. (C.58) with Eq. (C.56), we recover Eq. (C.55). □

We now use the element-wise bound on $B^{(a)}_{zz'}$ from Lemma C.8 to apply Lemma C.9 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 C.10. *Under the same conditions as Lemma C.8, the minimal eigenvalue $\lambda_1^{(a)}(T)$ of the empirical inverse covariance matrix $\frac{1}{T}B^{(a)}(T)$ satisfies the lower bound*

$$\lambda_1^{(a)}(T) \geq \lambda_{\min}^{(a)}(T)/\tilde{\kappa}, \tag{C.59}$$

where $\lambda_{\min}^{(a)}(T)$ is the minimal eigenvalue of $\bar{B}^{(a)}(T)$ defined in Eq. (C.39)*, with probability at least $1-\delta_\lambda$, where*

$$\delta_\lambda := \frac{Z^3}{(\lambda_{\min}^{(a)}(T))^2}\left(\pi_{\min} - \tilde{\kappa}^{-1}\right)^{-2}\frac{2}{T\gamma_\phi}(\kappa + \log\log(1/\rho_{\min})), \tag{C.60}$$

for any $\tilde{\kappa} \in (1/\pi_{\min}, \tilde{\kappa}_{\max})$, with

$$\frac{1}{\tilde{\kappa}_{\max}} = \pi_{\min} - \frac{Z}{\lambda_{\min}^{(a)}(T)}\sqrt{\frac{2}{T\gamma_\phi}(\kappa + \log\log(1/\rho_{\min}))}. \tag{C.61}$$

Proof. Recalling Eq. (C.54), we apply Lemma C.9 with

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

and have

$$\lambda_1^{(a)}(T) \geq \bar{\lambda}_1^{(a)}(T) - ZU_\delta, \tag{C.62}$$

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 C.6) that $\frac{1}{T}\mathbb{E}[B^{(a)}(T)] \succcurlyeq \pi_{\min}\bar{B}^{(a)}(T)$, or equivalently $\frac{1}{T}\mathbb{E}[B^{(a)}(T)] = \pi_{\min}\bar{B}^{(a)}(T) +$ PSD where PSD is a positive semidefinite symmetric matrix with non-negative minimal eigenvalue, and applying the Weyl inequality again (as in Lemma C.9), we have $\bar{\lambda}_1^{(a)}(T) \geq \pi_{\min}\lambda_{\min}^{(a)}(T)$, and thus,

$$\lambda_1^{(a)}(T) \geq \pi_{\min}\lambda_{\min}^{(a)}(T) - ZU_\delta. \tag{C.63}$$

Defining

$$\tilde{\kappa}^{-1} := \pi_{\min} - \frac{Z}{\lambda_{\min}^{(a)}(T)}\sqrt{\frac{1}{\delta}\frac{2}{T\gamma_\phi}(\kappa + \log\log(1/\rho_{\min}))}, \tag{C.64}$$

Eq. (C.63) takes the form of Eq. (C.59), with $\tilde{\kappa}$ inheriting its range, as stated in Lemma C.10 above, from the range of $\delta \in (0,1)$. Inverting Eq. (C.64) to express the probability $\delta_\lambda := Z\delta$ in terms of other parameters, we recover Eq. (C.60). □

C.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 C.5, 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. (C.18) – 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 $1-\delta$:

$$|\hat{\mu}_z^{(a)} - (\mu_\star^{(a)})_z| \leq \sum_{z'}|((B^{(a)})^{-1})_{zz'}| \cdot |g_{z'}^{(a)}| \leq T \cdot U_{\delta/Z}\sum_{z'}\left|((B^{(a)})^{-1})_{zz'}\right|. \tag{C.65}$$

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. (C.65) above, and recalling Lemma C.10, we have

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

with probability at least

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

With the definition of δ_λ in Eq. (C.60), recalling that U_δ^2 refers to the upper limit in Eq. (C.18), and setting $\tilde{\kappa} = 2/\pi_{\min}$ for simplicity, we recover Theorem 1 as stated above. □

Note from Eq. (C.60) (with $\tilde{\kappa} = 2/\pi_{\min}$) 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,

$$T\gamma_\phi > \frac{8Z^3}{\pi_{\min}\lambda_{\min}^{(a)}}(\kappa + \log\log(1/\rho_{\min})). \tag{C.66}$$

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.

D DERIVATION OF THEOREM 2

As outlined in the main text, the derivation of Theorem 2 involves (i) a generic procedure for converting bounds on empirical estimates into a regret bound for linear Thompson sampling, and (ii) application of Theorem 1 and related results, which bound empirical reward estimates and empirical covariance matrices in the latent bandit setting, to apply the resulting linear Thompson sampling regret bound to the latent bandit setting. This involves the following steps:

- In Appendix D.1, we define an important feature of the distribution over contexts in the linear bandit setting, which quantifies the amount of probability mass concentrated on contexts c_t where the reward gap between the optimal action $\text{argmax}_a c_t^\top \mu_\star^{(a)}$ and the next-best action is very small.
- In Appendix D.2 we state several assumptions used in our derivation, including bounds on empirical estimates which we later show to take a specific form in the case of Lemma 1 and Theorem 1 (where the linear bandit problem is obtained by reducing from the latent bandit setting, and conditioning on the true parameters $(\theta^\star, \phi^\star)$).
- In Appendix D.3, we derive (under these assumptions) a high-probability bound on the probability that linear Thompson sampling will select a suboptimal action at any time, given an observed context vector $c^\star$.
- In Appendix D.4 we upper bound (with high probability) the regret incurred by linear Thompson sampling at a given time, by taking an average over possible context vectors $c^\star$ of the mean regret incurred conditional on $c^\star$ (which is determined by the probability of suboptimal actions). We show that the bulk of expected regret comes from contexts $c^\star$ for which the best two actions have very similar expected reward.
- In Appendix D.5 we sum over timesteps to bound the cumulative regret of linear Thompson sampling. Since the regret bound at each timestep derived in Appendix D.4 fails with small but nonzero probability, an additional worst-case regret is incurred on timesteps when the bound fails. We optimize the probability of failure (a free parameter at each timestep) in order to tighten the bound on cumulative regret. Lastly, we specify from a more general case to the specific case in which empirical estimate error decreases as $1/\sqrt{T}$, which leads to $O(\sqrt{T})$ regret.
- Lastly, in Appendix D.6, we use Theorem 1, which bounds the error in empirical reward parameter estimators in the latent bandit setting (Section 3.1), along with a corresponding bound on empirical covariance matrices (Lemma C.10), to apply the generic linear TS regret bound of Appendix D.5 to the setting in which context vectors are posterior probability vectors over a latent state undergoing Markovian state transitions, $c_t^\star = p_t^\star$.

We remind the reader of the linear Thompson sampling algorithm (used by L^{2}TS as a subroutine), which is the focus of our analysis in this Appendix:

```
Linear Thompson Sampling (Agrawal and Goyal [2013b])
Input:
$\lambda_\mu > 0,\ \tilde{\sigma}_r^{(a)} > 0$ for $a \in \mathcal{A}$
$\hat{\mu}^{(a)} = \mathbf{0}_d$, $f^{(a)} = \mathbf{0}_d$, $B^{(a)} = \lambda_\mu \mathbb{1}_d$, for $a \in \mathcal{A}$
for $t \leftarrow 1, 2, ...$ do
Receive context $\hat{c}_t$
Sample $\mu^{(a)} \sim \mathcal{N}(\hat{\mu}^{(a)}, (\tilde{\sigma}_r^{(a)})^2 (B^{(a)})^{-1})$ for $a \in \mathcal{A}$
Select action $a = \text{argmax}_{a'} \hat{c}_t^\top \mu^{(a')}$
Observe reward $r_t$
Update mean reward estimates:
$B^{(a)} \leftarrow B^{(a)} + \hat{c}_t \hat{c}_t^\top,\ \ f^{(a)} \leftarrow f^{(a)} + \hat{c}_t r_t$
$\hat{\mu}^{(a)} = (B^{(a)})^{-1} f^{(a)}$
```

Preliminaries. We will distinguish the true context $c^\star$ – which determines the ground-truth mean reward, $\mathbb{E}[r_t | a_t = a] = (c_t^\star)^\top \mu_\star^{(a)}$ – from the context $\hat{c}$ which is accessible to the linear Thompson sampling agent. Throughout our analysis of linear Thompson sampling, we allow for error or corruption of observed contexts, $\hat{c} \neq c^\star$, which we assume to satisfy a bound (Assumption D.5). While we ultimately set $\hat{c} = c^\star$ when applying our analysis to the latent bandit setting, our regret bound for linear Thompson sampling (Lemmas D.2 and D.3, and Corollary D.3.1) applies more generally.

With the exception of Appendix D.6, we assume context feature vectors are in a d-dimensional Euclidean space, $c^\star, \hat{c} \in \mathbb{R}^d$. In Appendix D.6 we specify to our particular setting of interest, where $c^\star = p^\star$ is a probability vector restricted to the $(d-1)$-dimensional simplex, and $d = Z$ is the latent state dimensionality.

We use $P_c^{(t)}$ to denote the true distribution over linear bandit context vectors at time t, that is, $c_t^\star \sim P_c^{(t)}(\cdot)$, keeping in mind that in the latent bandit setting, $P_c^{(t)}$ will become the distribution over posterior probability vectors with elements $p_t^\star(z) := p(z_t = z | x_{1:t}, \theta^\star, \phi^\star)$, with the context history $x_{1:t}$ being a random sequence generated from given ground-truth parameters $(\theta^\star, \phi^\star)$.

Regarding notation, we define $\hat{\Omega}^{(a)} := (B^{(a)})^{-1}$ as the empirical covariance matrices used by linear Thompson sampling, and will assume $\tilde{\sigma}_r = 1$ for simplicity. We use $\hat{\mu} := \{\hat{\mu}^{(a)}\}_{a=1}^K$, $\hat{\Omega} := \{\hat{\Omega}^{(a)}\}_{a=1}^K$ to collectively denote the set of action-wise estimators and action-wise covariance matrices.

D.1 LINEAR BANDIT CONTEXT DISTRIBUTION

In this section, we define an important task-relevant feature of the context distribution $P_c^{(t)}$, which quantifies the likelihood of encountering contexts for which the optimal action has only marginally higher expected reward than the next-best action. Such "adversarial" contexts make it hard to resolve the best action, and are likely to induce suboptimal actions. As we will see, these regions of context space contribute significantly to expected regret.

Recall that, in the linear bandit setting of Section 3.2, a given context vector context c_t determines an optimal action $a_t^\star := \text{argmax}_a (c_t)^\top \mu_\star^{(a)}$, conditional on the true reward parameters $\{\mu_\star^{(a)}\}_{a=1}^K$. Thus, the space of context vectors may be partitioned into regions of optimality which favor different actions. (Note that in the latent bandit setting, this amounts to partitioning the simplex of probability vectors over the latent state.)

We will see that the asymptotic regret of linear Thompson sampling is controlled by the density of the context distribution near the borders of these regions of optimality, in the following way. We first define, for any context vector $c \in \mathbb{R}^d$ and pair of actions $(a^\star, a)$, the component $c_{\|}^{(a^\star,a)} \in \mathbb{R}$ parallel and perpendicular to the reward gap direction $\mu_\star^{(a^\star)} - \mu_\star^{(a)}$, that is,

$$c_{\|}^{(a,a^\star)}(c) := c^\top (\mu_\star^{(a^\star)} - \mu_\star^{(a)}) / ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_2. \tag{D.1}$$

$$c_{\perp}^{(a,a^\star)}(c) := \Pi_{a,a^\star} c, \tag{D.2}$$

where the projection matrix

$$\Pi_{a,a^\star} = \mathbb{1} - \frac{(\mu_\star^{(a^\star)} - \mu_\star^{(a)})(\mu_\star^{(a^\star)} - \mu_\star^{(a)})^\top}{||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_2^2}$$

projects c onto the $(d-1)$-dimensional hyperplane orthogonal to the vector difference $\mu_\star^{(a^\star)} - \mu_\star^{(a)}$.

Equivalently to Eq. (D.1), the difference in expected reward between actions $a^\star$ and a depends (only) on the parallel component of c,

$$c^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) = c_{||}^{(a^\star,a)} \Delta_{a^\star,a}, \tag{D.3}$$

where

$$\Delta_{a^\star,a} := ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_2 \tag{D.4}$$

is the magnitude of the vector difference of reward parameters for actions $a^\star$ and a. Thus, given fixed reward parameters $\mu_\star = \{\mu_\star^{(a)}\}_{a=1}^K$, the marginal distribution over the parallel component $c_{||}^{(a^\star,a)}$ of the context determines the probability distribution over the difference in expected rewards between $a^\star$ and a. Its density at small $c_{||}^{(a^\star,a)}$ quantifies the probability for "adversarial" contexts for which the better action (between $a^\star$ and a) becomes impossible to resolve. We define this limit as

$$\rho_{a^\star,a}^{(t)} := \lim_{\epsilon\to 0+} \frac{1}{\epsilon}\mathbb{P}_{c\sim P_c^{(t)}}\left(a(c) = a^\star, c^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) < \epsilon||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_2\right). \tag{D.5}$$

This quantity is the probability density which the distribution $P_c^{(t)}$ assigns to contexts for which action $a^\star$ is optimal, but is only infinitesimally preferred to action a. Note that, since the inequality in Eq. (D.5) can be written as $c_{||}^{(a^\star,a)} < \epsilon$, $\rho_{a^\star,a}^{(t)}$ only depends on the direction of the vector difference $\mu_\star^{(a^\star)} - \mu_\star^{(a)}$ (which determines $c_{||}^{(a^\star,a)}$, see Eq. (D.1)), and not its magnitude $\Delta_{a^\star,a}$. As mentioned above, when we interpret the context vectors c_t as posterior probability vectors in the latent bandit setting, $P_c^{(t)}$ becomes a distribution over these posteriors, which depend on the history of observations $x_{1:t}$. In this setting, Eq. (D.5) can be rewritten as shown in Eq. (10) in the main text.

D.2 ASSUMPTIONS

In this section we state several assumptions used throughout the derivation of Theorem 2. First, we make assumptions pertaining to the boundedness of context vectors and reward parameter vectors:

Assumption D.1. *The Euclidean norm of any context vector $c^\star \sim P_c^{(t)}$ (at any time t) is strictly upper bounded,*

$$||c^\star||_2 \leq u_c.$$

Assumption D.1 is automatically satisfied with $u_c = 1$ when the contexts $c^\star$ are posterior probability vectors.

Assumption D.2. *For all a, the 1-norm of the true mean reward parameter vector $\mu_\star^{(a)}$ is upper bounded,*

$$||\mu_\star^{(a)}||_1 < u_\mu.$$

We also make three generic assumptions on empirical estimates, which will be applied in Appendix D.3 to quantities at a single time t.

Assumption D.3. *For each (a, z) and for $\delta_1 \in (0,1)$, the error of the z'th vector element of the estimator $\hat{\mu}^{(a)}$ is upper bounded,*

$$|\hat{\mu}_z^{(a)} - (\mu_\star^{(a)})_z| \leq U_{\delta_1}^{(\hat{\mu})},$$

with probability at least $1-\delta_1$.

Assumption D.4. *For each a and for $\delta_2 \in (0,1)$, the maximal eigenvalue of the empirical covariance matrix $\hat{\Omega}^{(a)}$ is upper bounded,*

$$\max_z \lambda_z^{(a)} \leq (U_{\delta_2}^{(\hat{\Omega})})^2,$$

with probability at least $1-\delta_2$.

Assumption D.5. *The* 1*-norm error of the estimated context vector is upper bounded,*

$$||\hat{c} - c^\star||_1 := \sum_z |\hat{c}_z - c^\star_z| \leq U^{(\hat{c})}_{\delta_3},$$

with probability at least $1 - \delta_3$.

In general, we allow the upper bounds in Assumptions D.3-D.5 to be unspecified functions of the bound probabilities δ_i. In Appendix D.6, we will the use specific functional forms for these upper bounds (including time-dependence) which apply in the latent bandit setting under the conditions of Theorem 1. As noted above, in the latent bandit setting we will only consider the $U^{(\hat{c})}_{\delta_3} = 0$ case, but our analysis of linear Thompson sampling applies more generally.

D.3 UPPER BOUND ON SUBOPTIMAL ACTION PROBABILITIES

In this section, we show that for linear Thompson sampling, the probability of making a suboptimal action, given a Thompson sampling distribution $\mathcal{N}(\hat{\mu}, \hat{\Omega})$, can be bounded in terms of upper bounds on the error in $\hat{\mu}$ and on the eigenvalue spectrum of $\hat{\Omega}$, with suboptimal actions becoming impossible when the confidence ellipsoid determined by $\hat{\Omega}$ shrinks to zero (i.e. $U^{(\hat{\Omega})}_{\delta_2} \to 0$ in Assumption D.4) and reward error approaches zero.

The action probability bound, Eq. (D.7), is expressed in terms of a free parameter y which we will optimize in Appendix D.4 in order to tighten the resulting regret bound. Note that in Eq. (D.7), in the limit where u_1 and u_2 (which are proportional to the upper bounds in Assumptions D.3-D.5 on estimation error and uncertainty) become very small, y may be chosen to be very large, such that the probability of suboptimal action a can be upper bounded at a very small value except for contexts for which the reward gap $\Delta_a(c^\star)$ is infinitesimally small.

Lemma D.1. *When assumptions D.1 and D.3-D.5 are satisfied, the probability*

$$\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega}) := P(a_t = a|\hat{c}_t = \hat{c}, \hat{\mu}(t) = \hat{\mu}, \hat{\Omega}(t) = \hat{\Omega}) \tag{D.6}$$

of linear Thompson sampling selecting any action a *at any time* t*, conditional on empirical quantities* $(\hat{c}, \hat{\mu}, \hat{\Omega})$ *(the estimated or noisily observed context vector, the estimated reward parameters, and the empirical covariance matrix), with* $\hat{c} - c^\star$ *bounded by Assumption D.5, satisfies the upper bound*

$$\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega}) \leq \mathbf{1}\big(\Delta_a(c^\star) < yu_1 + u_2\big) + \frac{1}{2y} e^{-y^2} \tag{D.7}$$

for any $y > 0$*, with probability at least* $1 - \delta_3 - 2(d\delta_1 + \delta_2)$*, where* $c^\star$ *is the true context vector (whose difference from* $\hat{c}$ *is bounded by Assumption D.5),*

$$\Delta_a(c^\star) := (c^\star)^\top(\mu_\star^{(a(c^\star))} - \mu_\star^{(a)}), \tag{D.8}$$

is the context-dependent reward gap between action a *and the optimal action for context* $c^\star$*, and*

$$u_1 := 2\big(u_c + U^{(\hat{c})}_{\delta_3}\big)U^{(\hat{\Omega})}_{\delta_2}, \tag{D.9}$$

$$u_2 := U^{(\hat{c})}_{\delta_3}||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1 + 2dU^{(\hat{\mu})}_{\delta_1}(u_c + U^{(\hat{c})}_{\delta_3}). \tag{D.10}$$

Proof. For Thompson sampling, the action probabilities are averages over the multivariate normal distributions[7] from which the action-wise reward parameters $\mu^{(a)}$ are sampled:[8]

$$\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega}) = \int \prod_{a'} d\mu^{(a')} P_G(\mu^{(a')}|\hat{\mu}^{(a')}, \hat{\Omega}^{(a')}) \cdot \prod_{a' \neq a} \mathbf{1}(\hat{c}^\top \mu^{(a)} - \hat{c}^\top \mu^{(a')} > 0). \tag{D.11}$$

[7] We denote the multivariate Gaussian probability distribution function with mean μ and covariance Ω as $P_G(\cdot|\mu, \Omega)$.

[8] The second product over actions ensures that the probability for selecting action a is the integrated probability mass in the space of samples $\{\mu^{(a)}\}_{a=1}^K$ for which action a has the highest expected reward $\hat{c}^\top \mu_\star^{(a)}$.

For any $a^\star \neq a$, we can replace the indicator functions for all $a' \neq a, a^\star$ with 1, resulting in the upper bound:

$$\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega}) \leq \int d\mu^{(a)} P_G(\mu^{(a)}|\hat{\mu}^{(a)}, \hat{\Omega}^{(a)}) d\mu^{(a^\star)} P_G(\mu^{(a^\star)}|\hat{\mu}^{(a^\star)}, \hat{\Omega}^{(a^\star)}) \\ \cdot \mathbf{1}(\hat{c}^\top \mu^{(a)} - \hat{c}^\top \mu^{(a^\star)} > 0). \tag{D.12}$$

We define the difference in sampled reward parameters for actions a and $a^\star$ (which will be set to the optimal action $a(c^\star)$), shifted relative to the mean reward parameter estimators $(\hat{\mu}^{(a)}, \hat{\mu}^{(a^\star)})$, as

$$\nu := (\mu^{(a)} - \hat{\mu}^{(a)}) - (\mu^{(a^\star)} - \hat{\mu}^{(a^\star)}), \tag{D.13}$$

(For simplicity, we will suppress the implicit $(a, a^\star)$-dependence of ν.) Since the indicator function in Eq. (D.12) depends only on the difference

$$\mu^{(a)} - \mu^{(a^\star)} = \nu + \hat{\mu}^{(a)} - \hat{\mu}^{(a^\star)},$$

we can change variables from $(\mu^{(a)}, \mu^{(a^\star)})$ to $(\mu^{(a)} - \mu^{(a^\star)}, \mu^{(a)} + \mu^{(a^\star)})$ and integrate out the latter sum variable. The distribution of the difference $\mu^{(a)} - \mu^{(a^\star)}$ of two variables $\mu^{(a)} \sim P_G(\cdot|\hat{\mu}^{(a)}, \hat{\Omega}^{(a)})$ and $\mu^{(a^\star)} \sim P_G(\cdot|\hat{\mu}^{(a^\star)}, \hat{\Omega}^{(a^\star)})$ is Gaussian distributed with a covariance given by the sum of the individual covariances, that is,

$$\mu^{(a)} - \mu^{(a^\star)} \sim P_G(\cdot|\hat{\mu}^{(a)} - \hat{\mu}^{(a^\star)}, \hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)}). \tag{D.14}$$

In terms of the zero-mean variable ν, then, we can rewrite Eq. (D.12) – for any $a^\star \neq a$ – as

$$\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega}) \leq \int d\nu P_G(\nu|\mathbf{0}, \hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)}) \cdot \mathbf{1}(\hat{c}^\top \nu > \hat{c}^\top (\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})). \tag{D.15}$$

We now introduce a free parameter $\epsilon > 0$ and insert a factor of

$$1 = \mathbf{1}(\hat{c}^\top \nu \geq \epsilon ||\hat{c}||_2) + \mathbf{1}(\hat{c}^\top \nu < \epsilon ||\hat{c}||_2)$$

inside the integral in Eq. (D.15). This divides the space of samples ν into samples which are more or less optimistic about action a relative to $a^\star$ (relative to the estimated difference $\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)}$). Thus, for any $a^\star \neq a$,

$$\begin{aligned} \pi(a|\hat{c}, \hat{\mu}, \hat{\Omega}) &\leq \int d\nu P_G(\nu|\mathbf{0}, \hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)}) \cdot \mathbf{1}(\hat{c}^\top \nu > \hat{c}^\top (\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})) \\ &\qquad \times \big(\mathbf{1}(\hat{c}^\top \nu \geq \epsilon ||\hat{c}||_2) + \mathbf{1}(\hat{c}^\top \nu < \epsilon ||\hat{c}||_2)\big) \\ &\leq \mathbb{P}\Big(\hat{c}^\top \nu \geq \epsilon ||\hat{c}||_2 \Big| \nu \sim \mathcal{N}(\mathbf{0}, \hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)})\Big) + \mathbf{1}(\epsilon ||\hat{c}||_2 > \hat{c}^\top (\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})) \end{aligned} \tag{D.16}$$

In the first term we've used $\mathbf{1}(\hat{c}^\top \nu > \hat{c}^\top (\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})) \leq 1$, and in the second term we've upper bounded the indicator function,

$$\mathbf{1}(\hat{c}^\top \nu > \hat{c}^\top (\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})) \leq \mathbf{1}(\epsilon ||\hat{c}||_2 > \hat{c}^\top (\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})).$$

and taken the upper bound on the indicator function outside the integral, which is upper bounded by 1.

We will now use Assumption D.4 to derive an upper bounds on the first term in Eq. (D.16), and Assumptions D.3 and D.5 to correspondingly bound the second term.

Upper bound on the first term in Eq. (D.16).

Eq. (D.58) from Appendix D.7 gives an upper bound on the probability mass in the tail of a Gaussian distribution, which we use to upper bound the first term in (D.16). Recalling that the inner product of a Gaussian random vector $\nu \sim \mathcal{N}(\mathbf{0}, \Omega)$ with any vector c is a Gaussian variable with mean zero and variance $c^\top \Omega c$, Eq. (D.58) yields (for any $a^\star \neq a$)

$$\mathbb{P}\Big(\hat{c}^\top \nu \geq \epsilon ||\hat{c}||_2 \Big| \nu \sim \mathcal{N}(\mathbf{0}, \hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)})\Big) \leq \frac{1}{\sqrt{2}\epsilon} \sigma(\hat{c}) \exp\big[-\epsilon^2 / 2\sigma^2(\hat{c})\big], \tag{D.17}$$

where the variance

$$\sigma^2(\hat{c}) := ||\hat{c}||_2^{-2} \hat{c}^\top (\hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)}) \hat{c} \tag{D.18}$$

depends on the estimated context $\hat{c}$. To simplify the expectation over contexts required to compute regret, we upper bound $\sigma^2(\hat{c})$ in terms of the eigenvalues of the empirical covariance matrices $\{\hat{\Omega}^{(a)}\}$. Defining $\lambda_z^{(a,a^\star)}$ as the z'th eigenvalue of the covariance matrix $\hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)}$, the variance $\sigma^2(\hat{c})$ along any direction of the confidence ellipsoid specified by the same covariance matrix satisfies the upper bound

$$\sigma^2(\hat{c}) \leq \max_z \lambda_z^{(a,a^\star)}. \tag{D.19}$$

Furthermore, by the Weyl inequality (and since the matrices $\hat{\Omega}^{(a)}$ are real, symmetric, and positive definite),

$$\max_z \lambda_z^{(a,a^\star)} \leq \max_z \lambda_z^{(a)} + \max_z \lambda_z^{(a^\star)}.$$

Assumption D.4 states that $\max_z \lambda_z^{(a)},\ \max_z \lambda_z^{(a^\star)} \leq (U_{\delta_2}^{(\hat{\Omega})})^2$ with probability at least $(1-\delta_2)^2 > 1-2\delta_2$, and consequently,

$$\sigma^2(\hat{c}) \leq 2(U_{\delta_2}^{(\hat{\Omega})})^2 \tag{D.20}$$

with the same probability. Therefore, since the upper bound in Eq. (D.17) increases monotonically with $\sigma(\hat{c})$, we have, for any $\hat{c}$ and $a^\star \neq a$, and with probability at least $1 - 2\delta_2$,

$$\mathbb{P}\Big(\hat{c}^\top \nu \geq \epsilon ||\hat{c}||_2 \Big| \nu \sim \mathcal{N}(\mathbf{0}, \hat{\Omega}^{(a)} + \hat{\Omega}^{(a^\star)})\Big) \leq \frac{U_{\delta_2}^{(\hat{\Omega})}}{\epsilon} \exp\big[-\epsilon^2/(2U_{\delta_2}^{(\hat{\Omega})})^2\big] = \frac{1}{2y} e^{-y^2} \tag{D.21}$$

where

$$y := \frac{\epsilon}{2U_{\delta_2}^{(\hat{\Omega})}}, \tag{D.22}$$

is a rescaled version of the free parameter ϵ.

Upper bound on the second term in Eq. (D.16).

Applying Assumption D.3 to each element of the vector estimator $\hat{\mu}^{(a)}$, we have $||\hat{\mu}^{(a)} - \mu_\star^{(a)}||_1 \leq d \cdot U_{\delta_1}^{(\hat{\mu})}$ with probability at least $(1-\delta_1)^d > 1 - d\delta_1$. Applying this bound for both action a and $a^\star$, we have

$$\begin{aligned}\hat{c}^\top(\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)}) &= \hat{c}^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) + \hat{c}^\top(\hat{\mu}^{(a^\star)} - \mu_\star^{(a^\star)}) + \hat{c}^\top(\mu_\star^{(a)} - \hat{\mu}^{(a)}) \\ &\geq \hat{c}^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) - ||\hat{c}||_1 ||\hat{\mu}^{(a^\star)} - \mu_\star^{(a^\star)}||_1 - ||\hat{c}||_1 ||\hat{\mu}^{(a)} - \mu_\star^{(a)}||_1 \\ &\geq \hat{c}^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) - 2d||\hat{c}||_1 U_{\delta_1}^{(\hat{\mu})} \end{aligned} \tag{D.23}$$

with probability at least $(1-d\delta_1)^2 \geq 1 - 2d\delta_1$ (for any $a, a^\star \neq a$). It follows that

$$\mathbf{1}(\epsilon ||\hat{c}||_2 > \hat{c}^\top(\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})) \leq \mathbf{1}(\epsilon ||\hat{c}||_2 > \hat{c}^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) - 2d||\hat{c}||_1 U_{\delta_1}^{(\hat{\mu})}) \tag{D.24}$$

with the same probability. We will now apply Assumption D.5 to bound the deviation of the estimated context $\hat{c}$ from the true context $c^\star$. Recalling the shorthand notation $\Delta_a(c^\star) := (c^\star)^\top(\mu_\star^{(a(c^\star))} - \mu_\star^{(a)})$ of Eq. (D.8), and recalling that $a^\star = a(c^\star)$ is enforced in Eq. (D.33), we have

$$\hat{c}^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) = \Delta_a(c^\star) + (\hat{c} - c^\star)^\top (\mu_\star^{(a^\star)} - \mu_\star^{(a)}).$$

Applying Assumption D.5 to bound the error $\hat{c} - c^\star$, we have

$$|(\hat{c} - c^\star)^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)})| \leq ||\hat{c} - c^\star||_1 ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1 \leq U_{\delta_3}^{(\hat{c})} ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1$$

with probability at least $1 - \delta_3$. Consequently, $\hat{c}^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)}) \geq \Delta_a(c^\star) - U_{\delta_3}^{(\hat{c})} ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1$ with the same probability, and thus

$$\begin{aligned}&\mathbf{1}(\epsilon ||\hat{c}||_2 > \hat{c}^\top\big(\mu_\star^{(a^\star)} - \mu_\star^{(a)}\big) - 2d||\hat{c}||_1 U_{\delta_1}^{(\hat{\mu})}\big) \\ &\leq \mathbf{1}\big(\epsilon ||\hat{c}||_2 > \Delta_a(c^\star) - U_{\delta_3}^{(\hat{c})} ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1 - 2d||\hat{c}||_1 U_{\delta_1}^{(\hat{\mu})}\big) \\ &\leq \mathbf{1}\big(\Delta_a(c^\star) < \epsilon(||c^\star||_1 + U_{\delta_3}^{(\hat{c})}) + U_{\delta_3}^{(\hat{c})} ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1 + 2dU_{\delta_1}^{(\hat{\mu})}(||c^\star||_1 + U_{\delta_3}^{(\hat{c})})\big), \end{aligned} \tag{D.25}$$

with probability at least $1 - \delta_3$. In the last line, we have again used Assumption D.4 to exchange $\hat{c}$ in favor of $c^\star$, by using

$$||\hat{c}||_2 \leq ||\hat{c}||_1 = ||c^\star + (\hat{c} - c^\star)||_1 \leq ||c^\star||_1 + ||\hat{c} - c^\star||_1 \leq ||c^\star||_1 + U_{\delta_3}^{(\hat{c})}$$

in the first term, and similarly $||\hat{c}||_1 \leq ||c^\star||_1 + U_{\delta_3}^{(\hat{c})}$ in the last term. Lastly, we use Assumption D.1 to upper bound $||c^\star||_1$ in Eq. (D.25). Combining Eq. (D.25) with Eq. (D.24), we then have

$$\mathbf{1}(\epsilon||\hat{c}||_2 > \hat{c}^\top(\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})) \leq \mathbf{1}\big(\Delta_a(c^\star) < \epsilon(u_c + U_{\delta_3}^{(\hat{c})}) + U_{\delta_3}^{(\hat{c})}||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1 + 2dU_{\delta_1}^{(\hat{\mu})}(u_c + U_{\delta_3}^{(\hat{c})})\big), \tag{D.26}$$

with probability at least $(1-\delta_3)(1-2d\delta_1) \geq 1 - \delta_3 - 2d\delta_1$. To simplify the expression, we introduce the variables u_1 and u_2 defined above in Eqs. (D.9)-(D.10), which summarize the influence of the error bounds from Assumptions D.3-D.5, and write Eq. (D.26) as

$$\mathbf{1}(\epsilon||\hat{c}||_2 > \hat{c}^\top(\hat{\mu}^{(a^\star)} - \hat{\mu}^{(a)})) \leq \mathbf{1}(\Delta_a(c^\star) < yu_1 + u_2), \tag{D.27}$$

again with probability at least $1 - \delta_3 - 2d\delta_1$, where y was defined in Eq. (D.22). Finally, combining Eqs. (D.21) and (D.27) to upper bound (respectively) the first and second terms in Eq. (D.16), we arrive at the final high-probability bound on action probabilities, Eq. (D.7), with the probability of the bound obtained by combining the probabilities of Eqs. (D.21) and (D.27). □

D.4 INSTANTANEOUS REGRET BOUND

The suboptimal action probability bound, Lemma D.1, conditions on a particular context vector $\hat{c}$, which is approximately equal to the true context $c^\star$ (with the difference bounded by Assumption D.5). We now take an expectation over the context distribution $P_c^{(t)}$ from which $c^\star$ is generated at time t, in order to extend Lemma D.1 into a corresponding high-probability bound on the expected regret incurred at time t.

Lemma D.2. *When Assumptions D.1 and D.3-D.5 are satisfied, and furthermore when*

$$(U_{\delta_2}^{(\hat{\Omega})})^2 < \frac{1}{8e}\frac{\Delta_{a^\star,a}^2}{u_c\rho_{a^\star,a}^{(t)}} \tag{D.28}$$

the expected regret incurred by linear Thompson sampling at a single timestep is upper bounded,

$$\delta\mathcal{R}^{(t)} \leq \sum_{a^\star,a}\frac{\rho_{a^\star,a}^{(t)}}{\Delta_{a^\star,a}}\Big(\big(||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_1 U_{\delta_3}^{(\hat{c})} + 2du_cU_{\delta_1}^{(\hat{\mu})}\big)^2 + 8u_c^2\big(U_{\delta_2}^{(\hat{\Omega})}\big)^2\log\zeta\Big) + O(U^3) \tag{D.29}$$

with probability at least $1 - 2(d\delta_1 + \delta_2) - \delta_3$, *where*

$$\zeta := \frac{\Delta_{a^\star,a}^2}{2\rho_{a^\star,a}^{(t)}}\frac{1}{u_c(2U_{\delta_2}^{(\hat{\Omega})})^2}, \tag{D.30}$$

and where $O(U^3)$ *denotes contributions which scale cubically or higher with the upper bounds* $U_{\delta_1}^{(\hat{\mu})}$, $U_{\delta_2}^{(\hat{\Omega})}$, $U_{\delta_3}^{(\hat{c})}$ *on estimation errors.*

Proof. The instantaneous or per-timestep expected regret incurred by selecting action a_t – averaged over possible ground-truth context vectors $c^\star$ (and $\hat{c} \approx c^\star$ up to error bounded by Assumption D.5) and actions a, but conditioned on the empirical estimates $(\hat{\mu}, \hat{\Omega})$ – is

$$\delta\mathcal{R}^{(t)}(\hat{\mu}, \hat{\Omega}) = \sum_a \mathbb{E}_{c^\star\sim P_c^{(t)}}[(c^\star)^\top(\mu_\star^{(a(c^\star))} - \mu_\star^{(a)})\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega})] \tag{D.31}$$

$$= \sum_{a^\star,a}\delta\mathcal{R}_{a^\star,a}^{(t)}(\hat{c}, \hat{\mu}, \hat{\Omega}), \tag{D.32}$$

where

$$\delta\mathcal{R}_{a^\star,a}^{(t)}(\hat{\mu}, \hat{\Omega}) := \mathbb{E}_{c^\star\sim P_c^{(t)}}\big[\mathbf{1}(a(c^\star) = a^\star)\Delta_a(c^\star)\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega})\big] \tag{D.33}$$

is the pair-wise expected regret incurred due to taking action a when $a^\star$ is optimal, and we have used the definition of the reward gap $\Delta_a(c)$ in Eq. (D.8). Using Lemma D.1 to upper bound the action probability $\pi(a|\hat{c}, \hat{\mu}, \hat{\Omega})$ in Eq. (D.33), the action pair-wise regret $\delta\mathcal{R}_{a^\star,a}^{(t)}$ satisfies the upper bound

$$\delta\mathcal{R}_{a^\star,a}^{(t)} \leq \mathbb{E}_{c^\star\sim P_c^{(t)}}\Big[\mathbf{1}(a(c^\star) = a^\star)\Delta_a(c^\star)\Big(\mathbf{1}(\Delta_a(c^\star) < yu_1 + u_2) + \frac{1}{2y}e^{-y^2}\Big)\Big] \tag{D.34}$$

for any $y > 0$, with probability at least $1 - \delta_{\mathcal{R}}$, where

$$\delta_{\mathcal{R}} := 2(d\delta_1 + \delta_2) + \delta_3. \tag{D.35}$$

We have removed the arguments $(\hat{\mu}, \hat{\Omega})$ of regret, since the upper bound holds for any values of these arguments, and thus also bounds the expected regret, averaged over possible realizations of these estimators, $\delta\mathcal{R}^{(t)}_{a^\star,a} = \mathbb{E}_{\hat{\mu},\hat{\Omega}}[\delta\mathcal{R}^{(t)}_{a^\star,a}(\hat{\mu}, \hat{\Omega})]$.

Asymptotic limit of small errors.

In the asymptotic, large-T limit, we expect that the u_1 and u_2 – which scale linearly with the upper bounds on the errors $(U^{(\hat{\mu})}_{\delta_1}, U^{(\hat{\Omega})}_{\delta_2}, U^{(\hat{c})}_{\delta_3})$ in $(\hat{\mu}, \hat{\Omega}, \hat{c})$ in Assumptions D.3-D.5 – will converge towards zero. In this regime, the indicator function in the first term in Eq. (D.34) will only be nonzero when the context-dependent reward gap between actions a and $a^\star$ is very small, making it difficult to resolve the better action. Defining $\bar{\epsilon} := yu_1 + u_2$ for brevity, the first term in Eq. (D.34) can be evaluated as follows, by expressing the expectation over $c^\star$ as an integral over the parallel component $c_{\|}^{(a^\star,a)} \sim P_{\|}^{(t)}(\cdot|a^\star, a)$ introduced above in (D.1):

$$\begin{aligned}
&\mathbb{E}_{c\sim P_c^{(t)}}\Big[\mathbf{1}(a(c) = a^\star)\Delta_a(c) \cdot \mathbf{1}\big(\Delta_a(c) < \bar{\epsilon}\big)\Big] = \int_0^{\bar{\epsilon}/\Delta_{a^\star,a}} dx P_{\|}^{(t)}(x|a^\star, a) x \Delta_{a^\star,a} \mathbb{P}(a(c) = a^\star | c_{\|}^{(a^\star,a)} = x) \\
&= \Delta_{a^\star,a} \int_0^{\bar{\epsilon}/\Delta_{a^\star,a}} dx \cdot x \times \left[\lim_{x\to 0+} P_{\|}^{(t)}(x|a^\star, a) \cdot \mathbb{P}(a(c) = a^\star | c_{\|}^{(a^\star,a)} = x) + O(\bar{\epsilon})\right] \\
&= \frac{\bar{\epsilon}^2}{2\Delta_{a^\star,a}} \left[\frac{1}{\bar{\epsilon}} \int_0^{\bar{\epsilon}} dx P_{\|}^{(t)}(x|a^\star, a) \cdot \mathbb{P}(a(c) = a^\star | c_{\|}^{(a^\star,a)} = x) + O(\bar{\epsilon})\right] \\
&= \frac{\bar{\epsilon}^2}{2\Delta_{a^\star,a}} \left[\frac{1}{\bar{\epsilon}} \mathbb{P}_{c\sim P_c^{(t)}} \left(a(c) = a^\star, c_{\|}^{(a^\star,a)} < \bar{\epsilon}\right) + O(\bar{\epsilon})\right]
\end{aligned} \tag{D.36}$$

In the first line, we have conditioned on the event $(a(c^\star) = a^\star)$ that $a^\star$ is optimal by restricting $x \sim P_{\|}^{(t)}(\cdot|a^\star, a)$ to be positive,[9] In the second line, since $x < \bar{\epsilon}/\Delta_{a^\star,a}$, we have written the integrant as its limit as the parallel component approaches zero, up to $O(\bar{\epsilon})$ corrections. In the third line, we have evaluated the integral over x and rewritten the limiting quantity in brackets as an integral, which is exact up to an additional $O(\bar{\epsilon})$ correction. In the last line, we have rewritten the integral over the marginal and joint distributions as a joint probability. In the $\bar{\epsilon} \to 0$ limit, recalling the definition of $c_{\|}^{(a^\star,a)}$ in Eqs. (D.1)-(D.3), this final quantity in brackets is the limiting pairwise probability density $\rho_{a^\star,a}$ defined above in Eq. (D.5). Therefore, up to an additional $O(\bar{\epsilon})$ correction due to the change in this quantity away from its limit as $\bar{\epsilon} \to 0$, we have

$$\mathbb{E}_{c\sim P_c^{(t)}}\Big[\mathbf{1}(a(c) = a^\star)\Delta_a(c) \cdot \mathbf{1}\big(\Delta_a(c) < \bar{\epsilon}\big)\Big] = \frac{\rho^{(t)}_{a^\star,a}}{2\Delta_{a^\star,a}} \bar{\epsilon}^2 + O(\bar{\epsilon}^3) \tag{D.37}$$

While this limiting form obviously fails for large $\bar{\epsilon}$, at late times we expect the error bounds $(U^{(\hat{\mu})}_{\delta_1}, U^{(\hat{\Omega})}_{\delta_2}, U^{(\hat{c})}_{\delta_3})$ to become tight, and hence $\bar{\epsilon}$ to approach zero.

The expectation in the second term in Eq. (D.34) can also be upper bounded,

$$\mathbb{E}_{c^\star\sim P_c^{(t)}}[\mathbf{1}(a(c^\star) = a^\star)\Delta_a(c^\star)] \le \Delta_{a^\star,a} \cdot \mathbb{E}_{c^\star\sim P_c^{(t)}}[||c^\star||_2] \le u_c \Delta_{a^\star,a}. \tag{D.38}$$

Here, we have used $\mathbf{1}(a(c^\star) = a^\star) \le 1$, set $a(c^\star) = a^\star$, upper bounded the vector inner product $\Delta_a(c^\star)$ in terms of the Euclidean norms $||c^\star||_2$ and $||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_2 = \Delta_{a^\star,a}$, used $||c^\star||_2 \le ||c^\star||_1$, and used Assumption D.1 again.

Applying Eq. (D.37) and Eq. (D.38) in Eq. (D.34), $\delta\mathcal{R}_{a^\star,a}$ can be upper bounded as follows:

$$\delta\mathcal{R}^{(t)}_{a^\star,a} \le \frac{\rho^{(t)}_{a^\star,a}}{2\Delta_{a^\star,a}} (yu_1 + u_2)^2 + u_c \Delta_{a^\star,a} \frac{1}{2y} e^{-y^2} + O(U^3), \tag{D.39}$$

with probability at least $1 - \delta_{\mathcal{R}}$ as specified in Eq. (D.35), where $O(U^3)$ indicates contributions which scale with the cube of the error bounds $(U^{(\hat{\mu})}_{\delta_1}, U^{(\hat{\Omega})}_{\delta_2}, U^{(\hat{c})}_{\delta_3})$. This is in contrast with the leading terms, which scale quadratically with u_1 and u_2

[9] Recall from Eq. (D.1) that the sign of $c_{\|}^{(a^\star,a)}$ specifies whether or not action $a^\star$ is preferred to a.

and hence[10] with the error bounds. In the limit where all errors become very small, the $O(U^3)$ contribution will become negligible compared to the leading terms.

Optimization of the free parameter y.

We are now in a position to optimize the free parameter y. Noting that

$$(yu_1 + u_2)^2 \leq 2(y^2 u_1^2 + u_2^2),$$

and defining

$$u := \frac{2\rho^{(t)}_{a^\star,a}}{\Delta^2_{a^\star,a} u_c} u_1^2, \quad v := \frac{1}{2} u_c \Delta_{a^\star,a} \tag{D.40}$$

to simplify notation, we have

$$\delta\mathcal{R}^{(t)}_{a^\star,a} \leq \frac{\rho^{(t)}_{a^\star,a}}{\Delta_{a^\star,a}} u_2^2 + v\left(uy^2 + y^{-1}e^{-y^2}\right) + O(U^3) \tag{D.41}$$

Defining a rescaled variable $\tilde{y} := ue^{y^2}$, the second term is

$$uv \times \Big(\log(\tilde{y}/u) + \frac{1}{\tilde{y}\sqrt{\log(\tilde{y}/u)}}\Big).$$

Setting $\tilde{y} = 1$ for simplicity, so that $y^2 = \log(1/u)$, it is straightforward to check that the first term is larger as long as $u < 1/e \approx 0.368$. Under this assumption and with this choice of y, then, the second term in Eq. (D.41) is $\leq 2uv\log(1/u)$, and hence

$$\delta\mathcal{R}^{(t)}_{a^\star,a} \leq \frac{\rho^{(t)}_{a^\star,a}}{\Delta_{a^\star,a}} \left(u_2^2 + 2u_1^2 \log\left(\frac{\Delta^2_{a^\star,a}}{2\rho^{(t)}_{a^\star,a}} \frac{u_c}{u_1^2}\right)\right) + O(U^3) \tag{D.42}$$

Finally, absorbing the terms in u_1 and u_2 – see Eqs. (D.9)-(D.10) – which scale quadratically with the error bounds $(U^{(\hat{\mu})}_{\delta_1}, U^{(\hat{\Omega})}_{\delta_2}, U^{(\hat{c})}_{\delta_3})$ into the subleading $O(U^3)$ contribution, and summing over actions as in Eq. (D.32), we arrive at the final form of the instantaneous regret bound, Eq. (D.29). (The condition that $u < 1/e$ is given in Eq. (D.28), again with the higher order term in u_1 removed.) □

D.5 REGRET BOUND FOR LINEAR THOMPSON SAMPLING

We now sum over timesteps in order to extend the per-timestep regret bound from the previous section into a cumulative regret bound. In the following Lemma, we assume a generic form for the per-timestep bound, which will be partially specified in the subsequence Corollary, and fully specified using Lemma D.2 above in the following section.

Lemma D.3. *When Assumption D.1 is satisfied (such that $||c^\star||_2 \leq u_c$), and when the per-timestep regret at any time t for a given algorithm satisfies the upper bound*

$$\delta\mathcal{R}^{(t)} \leq \frac{U_\mathcal{R}}{t^{\nu_1}\delta_t^{\nu_2}} \tag{D.43}$$

with probability at least $1 - \delta_t$, for any $\delta_t \in (0, 1)$ and for $\nu_1 \in (0, 1], \nu_2 \in [0, 1]$ with $1 - \nu_1/(1 + \nu_2) > 0$, the corresponding cumulative regret $\mathcal{R}(T) = \sum_{t=1}^T \delta\mathcal{R}^{(t)}$ satisfies the upper bound

$$\mathcal{R}(T) \leq U_\mathcal{R}^{1/(1+\nu_2)} (u_c \cdot \max_{a^\star,a} \Delta_{a^\star,a})^{\nu_2/(1+\nu_2)} \frac{1+\nu_2}{1+\nu_2-\nu_1} T^{1-\nu_1/(1+\nu_2)}. \tag{D.44}$$

Proof. At any given time, the per-timestep regret satisfies (with probability 1) the bound

$$\delta\mathcal{R}^{(t)} \leq \frac{U_\mathcal{R}}{t^{\nu_1}\delta_t^{\nu_2}} + \delta_t \times \max_{c^\star} \max_{a,a^\star} |(c^\star)^\top (\mu_\star^{(a^\star)} - \mu_\star^{(a)})| \tag{D.45}$$

[10] Below, we will set the free parameter y to an optimal value which scales only logarithmically with the error bounds.

for some $\nu_1, \nu_2 > 0$. The second term conservatively bounds the worst-case regret incurred when the high-probability bound, Eq. (D.43), fails with probability $\leq \delta_t$. We choose a power-law time schedule for the bound probability parameter δ_t,

$$\delta_t = \delta_0 / t^{\nu_\delta},$$

with free parameters $\delta_0 > 0$ and $\nu_\delta > 0$. With this schedule, and using Assumption D.1 to bound $|(c^\star)^\top(\mu_\star^{(a^\star)} - \mu_\star^{(a)})| \leq u_c ||\mu_\star^{(a^\star)} - \mu_\star^{(a)}||_2 = u_c \Delta_{a^\star,a}$, we have

$$\delta\mathcal{R}^{(t)} \leq \frac{U_\mathcal{R}}{\delta_0^{\nu_2}} t^{\nu_2 \nu_\delta - \nu_1} + \delta_0 t^{-\nu_\delta} u_c \cdot \max_{a^\star,a} \Delta_{a^\star,a}.$$

The sum over timesteps can be bounded with the continuous integral, $\sum_{t=1}^T t^{-\nu} \leq \int_0^T t^{-\nu} dt = \frac{1}{1-\nu} T^{1-\nu}$, as long as $\nu \in (0,1)$. Assuming, then, that $\nu_\delta \in (0,1)$, $\nu_1 - \nu_2 \nu_\delta \in (0,1)$, we have

$$\mathcal{R}(T) = \sum_{t=1}^T \delta\mathcal{R}^{(t)} \leq \frac{U_\mathcal{R}}{\delta_0^{\nu_2}(1 - (\nu_1 - \nu_2\nu_\delta))} T^{1-(\nu_1 - \nu_2\nu_\delta)} + \frac{\delta_0}{1-\nu_\delta} T^{1-\nu_\delta} u_c \cdot \max_{a^\star,a} \Delta_{a^\star,a}. \tag{D.46}$$

The free parameter ν_δ controls the tradeoff between the growth rates in time of the two terms. Equating these exponents,

$$1 - (\nu_1 - \nu_2 \nu_\delta) = 1 - \nu_\delta,$$

leads to $\nu_\delta = \nu_1/(\nu_2 + 1) \in (0,1)$. Consequently,

$$\mathcal{R}(T) = \sum_{t=1}^T \delta\mathcal{R}^{(t)} \leq \left(\frac{U_\mathcal{R}}{\delta_0^{\nu_2}} + \delta_0 u_c \cdot \max_{a^\star,a} \Delta_{a^\star,a} \right) \frac{1+\nu_2}{1+\nu_2-\nu_1} T^{1-\nu_1/(1+\nu_2)}. \tag{D.47}$$

The free parameter δ_0 can be optimized by setting its derivative to zero, which yields

$$\delta_0 = \left(\frac{U_\mathcal{R} \nu_2}{u_c \max_{a^\star,a} \Delta_{a^\star,a}} \right)^{1/(\nu_2+1)}.$$

Using this value of δ_0 in Eq. (D.47), along with the assumed condition that $\nu_2 \leq 1$, we recover Eq. (D.44). □

While Eq. (D.43) allows for a generic power-law time-dependence of the per-timestep regret bound and its probability of failure δ_t, in practice the exponents (ν_1, ν_2) will take specific values. In particular, in the limit of approximately i.i.d. reward data, the error $||\hat{\mu}^{(a)} - \mu_\star^{(a)}||$ in reward estimators decreases as $1/\sqrt{t}$, and can be bounded (e.g. as shown in Appendix C) for any reward distribution using Chebyshev's inequality, resulting in the following specific case of Lemma D.3:

Corollary D.3.1. *When Assumption D.1 is satisfied, and when the per-timestep regret at any time t for a given algorithm satisfies the upper bound*

$$\delta\mathcal{R}^{(t)} \leq \frac{U_\mathcal{R}}{t \cdot \delta_t} \tag{D.48}$$

with probability at least $1 - \delta_t$*, for any* $\delta_t \in (0,1)$*, the cumulative regret* $\mathcal{R}(T) = \sum_{t=1}^T \delta\mathcal{R}^{(t)}$ *satisfies the upper bound*

$$\mathcal{R}(T) \leq 2 \left(U_\mathcal{R} \cdot u_c \cdot \max_{a^\star,a} \Delta_{a^\star,a} \right)^{1/2} T^{1/2}. \tag{D.49}$$

Proof. Eq. (D.49) is the special case of Eq. (D.44) for $\nu_1 = \nu_2 = 1$. □

D.6 REGRET BOUND FOR LATENT LINEAR THOMPSON SAMPLING

We are now in a position to apply the cumulative regret bound of the previous section, along with the specific form of the per-timestep regret for linear Thompson sampling from Lemma D.2 and the latent bandit error bound, Theorem 1, to finally derive Theorem 2:

Proof of Theorem 2. In the latent bandit setting of Section 3.1, Theorem 1 guarantees that Assumption D.3 is satisfied, at time t, with

$$U_{\delta_1}^{(\hat{\mu})} = \frac{2Z^2}{\pi_{\min}^2 \cdot \min_a \lambda_{\min}^{(a)}(t)} \sqrt{\frac{1}{\delta_1 \cdot t}\Big(\sigma_{\text{eq}}^2 + \frac{4u_\mu^2}{\gamma_{\phi^\star}}(1+\log\zeta_{\phi^\star})\Big)} + O(1/t^{3/2}) \tag{D.50}$$

with probability at least $1-\delta_1$. Here, we have defined $1-\delta_1$ to be the probability of the bound as given in Eq. (7), and have Taylor expanded the $O(1/t)$ contribution in $\delta_1 = \delta + O(1/t)$ from Eq. (7) into a $O(1/t^{3/2})$ contribution to $U_{\delta_1}^{(\hat{\mu})}$. We have also used Assumption D.2 to bound $||\mu_\star^{(a)}||_1 < u_\mu$. (We remind the reader that the definitions of quantities in Eq. (D.50) are given in Theorem 1.)

Likewise, Lemma C.10 (which was used to derive Theorem 1) guarantees that Assumption D.4 is satisfied under the same conditions, at time t, with

$$(U_{\delta_2}^{(\hat{\Omega})})^2 = \frac{\tilde{\kappa}}{t\lambda_{\min}^{(a)}(t)} \tag{D.51}$$

with probability of failure $\delta_2 \propto (\pi_{\min} - \tilde{\kappa}^{-1})^{-2}/t$, as shown in Eq. (C.60). Here we have converted the minimal eigenvalue lower bound of Eq. (C.59) into a maximal eigenvalue upper bound for the inverse matrix, $\hat{\Omega}^{(a)}(t) = \frac{1}{t}(B^{(a)}(t))^{-1}$.

While $\lambda_{\min}^{(a)}(t)$ introduces time-dependence in Eqs. (D.50) and (D.51), this time-dependence can be ignored at late times, where all quantities converge to limiting asymptotic forms. Under the assumption of ergodicity of the latent Markov chain (used in Appendix C for Theorem 1), as $T \to \infty$ the latent state converges to an equilibrium distribution. Consequently, the generating distribution of context data $x_{t-\tau:t}$ approaches an asymptotic equilibrium distribution for any fixed τ, as $t \to \infty$. Since the posteriors probabilities $p_t^\star(z) = p(z_t = z|x_{1:t})$ are deterministic functions of context data, and furthermore since the dependence on past data $x_{t' \ll t}$ becomes exponentially suppressed with decay rate $\gamma_{\phi^\star}$ (see Appendix C.1), the distribution over these posterior probabilities will also converge exponentially quickly to an asymptotic equilibrium form at late times. Thus, setting $c_t^\star = p_t^\star$, we see that the distribution $P_c^{(t)}$ over linear bandit context vectors converges to an asymptotic distribution, with differences decaying exponentially in time with a decay rate $\gamma_{\phi^\star}$. Recall that the action-wise inverse covariance matrices

$$\bar{B}^{(a)}(T) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_{1:t}}\left[\mathbf{1}(a = a_t^\star) p_t^\star (p_t^\star)^\top\right],$$

defined in Eq. (C.39), are sums of expectations over the posteriors $p_t^\star$. As $T \to \infty$, the contributions from $t < \sqrt{T}$ will decrease as $O(\sqrt{T}/T) = O(1/\sqrt{T})$, and will thus make an $O(1/\sqrt{T})$ contribution to the minimal eigenvalues $\lambda_{\min}^{(a)}(T)$. This can be absorbed into the $O(t^{-3/2})$ late-time corrections in Eq. (D.50). Furthermore, contributions to $\bar{B}^{(a)}(T)$ from $t > \sqrt{T}$ will be expectations over the limiting equilibrium distribution over $p_t^\star$, up to differences decaying exponentially, which can also be absorbed into the $O(t^{-3/2})$ corrections in Eq. (D.50).[11] Therefore, from now on, we set the minimal eigenvalues in Eqs. (D.50) and (D.51) equal to their asymptotic values, which we define as

$$\lambda_{\min}^{(a)} = \lim_{t\to\infty} \lambda_{\min}^{(a)}(t). \tag{D.52}$$

Furthermore, we define $\lambda_{\min} := \min_a \lambda_{\min}^{(a)}$ as the minimum eigevalue over all action-wise inverse covariance matrices, $\bar{B}^{(a)}(t)$ as $t \to \infty$.

Lastly, under the assumption of Theorem 1 that the true posteriors $p_t^\star$ are used as context vectors, Assumption D.4 is satisfied with $U_{\delta_3}^{(\hat{c})} = 0$ and $\delta_3 = 0$.

Given these error bounds, and setting $d = Z$ and $u_c = 1$ (since $||p_t^\star||_2 \leq ||p_t^\star||_1 = \sum_z p_t^\star(z) = 1$), Lemma D.2 takes the form

$$\delta\mathcal{R}^{(t)} \leq \sum_{a^\star,a} \frac{\rho_{a^\star,a}^{(t)}}{\Delta_{a^\star,a}} \Big(\big(2ZU_{\delta_1}^{(\hat{\mu})}\big)^2 + 8\big(U_{\delta_2}^{(\hat{\Omega})}\big)^2 \log\zeta\Big) + O((\delta_1 t)^{-3/2}) \tag{D.53}$$

[11] This relies on a strictly positive decay rate $\gamma_{\phi^\star}$. However, when $\gamma_{\phi^\star}$, the upper bound $U_{\delta_1}^{(\hat{\mu})}$ becomes vacuous anyways.

with probability at least $1 - 2Z\delta_1 - \delta_2$, where $(U^{(\hat{\mu})}_{\delta_1}, U^{(\hat{\Omega})}_{\delta_2})$ are given in Eqs. (D.50) and (D.51), and we have set $U^{(\hat{c})}_{\delta_3} = 0$, and where ζ was defined in Eq. (D.30). Here, we have used the fact that $U^{(\hat{\mu})}_{\delta_1} \propto (1/\sqrt{t\delta_1})$ in Eq. (D.50) in the $O(U^3)$ contributions in Eq. (D.29). (Note that in the context of linear Thompson sampling, the minimal probability $\pi_{\min}$ of selecting the optimal action can be lower bounded at $1/K$ by initializing the empirical covariance matrices to allow for sufficiently broad posteriors over $\mu^{(a)}_\star$.)

The $U^{(\hat{\Omega})}_{\delta_2}$ term only contributes subleading corrections to regret, as $t \to \infty$, for the following reason. At times t when the bound on $U^{(\hat{\Omega})}_{\delta_2}$ holds with probability $1 - \delta_2$, the regret incurred scales as $1/t$. Additional regret is incurred from times when the bound on $U^{(\hat{\Omega})}_{\delta_2}$ fails. This occurs with probability $\delta_2 \propto 1/t$, yielding additional per-timestep expected regret that is also $O(1/t)$.

Thus, with the $U^{(\hat{\Omega})}_{\delta_2}$ term incorporated into the subleading corrections, and furthermore using Eq. (D.52) in the $t \to \infty$ limit (as discussed above), we now have

$$\delta\mathcal{R}^{(t)} \leq U \times \sum_{a^\star, a} \frac{\rho_{a^\star,a}}{\Delta_{a^\star,a}} \; + O((\delta_1 t)^{-3/2}) + O(1/t) + O(t^{-2}) \tag{D.54}$$

with probability[12] at least $1 - 2Z\delta_1$. where

$$U := \frac{16Z^6}{\pi^4_{\min}\lambda^2_{\min}} \frac{1}{\delta_1 \cdot t} \Big(\sigma^2_{\text{eq}} + \frac{4u^2_\mu}{\gamma_{\phi^\star}} \big(1 + \log \zeta_{\phi^\star}\big)\Big)$$

We have also omitted the time index on $\rho_{a^\star,a}$, which we define as the asymptotic limit

$$\rho_{a^\star,a} := \lim_{t\to\infty} \rho^{(t)}_{a^\star,a}. \tag{D.55}$$

This is because, since $\rho^{(t)}_{a^\star,a}$ is also an expectation over the current distribution $P^{(t)}_c$ of context vectors, it will converge to a fixed asymptotic value, with differences from its $t \to \infty$ limit decaying exponentially. As discussed above, these differences are smaller than the subleading corrections in Eq. (D.54), so we omit them.

In Eq. (D.54), the $O(t^{-2})$ contribution comes from the $O(t^{-3/2})$ contribution to $U^{(\hat{\mu})}_{\delta_1}$ in Eq. (D.50). Finally, defining the parameter δ_t in Eq. (D.48) as $\delta_t := 2Z\delta_1$, such that the per-timestep regret bound holds with probability $1 - \delta_t$, we can apply Corollary D.3.1. Plugging Eq. (D.54) into Eq. (D.48), Eq. (D.49) then recovers the final bound, Eq. (11).

Note that: (1) The $O(T^{2/5})$ scaling of the subleading corrections arises from applying Lemma D.3 to the $O((\delta_1 t)^{-3/2})$ contribution in Eq. (D.54), and setting $\nu_1 = \nu_2 = 3/2$ in Eqs. (D.43)- (D.44). (2) The $O(t^{-2})$ contributions to $\delta\mathcal{R}^{(t)}$ integrates to a constant when summing over t, which is (asymptotically) smaller than the $O(T^{2/5})$ correction. □

Problem-dependent structure of Theorem 2. We end this section by reminding the reader of the key dependencies in Theorem 2, described in the main text.

In addition to these dependencies, the Z-dependence and dependences of Δ_{likely} in Theorem 2 are inherited from Theorem 1, with the regret at time t being bounded proportional to the squared error, $||\hat{\mu}^{(a)} - \mu^{(a)}_\star||^2_2$. This dependence arises from the fact that increasing the error increases both (i) the size of the space of posterior beliefs $p^\star_t$ for which the true reward gap $(p^\star_t)^\top(\mu^{(a^\star)}_\star - \mu^{(a)}_\star)$ is too small to resolve relative to the error in estimating $\mu^{(a^\star)}_\star - \mu^{(a)}_\star$, which increases the probability of a suboptimal action, as well as (ii) the scale of the reward gap (regret incurred) when suboptimal actions are taken. In short, mistakes are made more frequently, and mistakes are more costly.

Furthermore, we note that regret is implicitly proportional to the number of actions K. This is because the inverse covariance $\bar{B}^{(a)}$ in Eq. (8) picks out only times when a given action a is optimal, and thus scales as $1/K$, becoming small when there are many actions to choose from. Consequently, the corresponding eigenvalues $\lambda^{(a)}_{\min}$ also scale as $1/K$, leading to regret proportional to K. This captures the fact that when there are many actions to explore, it takes longer to reduce uncertainty (bounded by $\lambda_{\min}$) about all of them.

[12] Recall that the probability δ_2 of the covariance bound failing can be chosen to decay to zero as $1/t$.

D.7 BOUND ON GAUSSIAN TAIL PROBABILITY MASS

The probability mass in the normal distribution $\mathcal{N}(0, \sigma)$ above threshold x is

$$\int_x^\infty dy \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} = \frac{1}{2}\left(1 - \mathrm{erf}(x/\sqrt{2}\sigma)\right), \tag{D.56}$$

where $\mathrm{erf}()$ is the error function, which can be expanded for large argument values and bounded,

$$\mathrm{erf}(z) > 1 - z^{-1} e^{-z^2} \tag{D.57}$$

for all $z > 0$, but tightly as $z \to \infty$. Equivalently, the probability mass in the tail is bounded as

$$\int_x^\infty dy \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} < \frac{\sigma}{\sqrt{2}x} e^{-x^2/2\sigma^2} \tag{D.58}$$

for $x > 0$.