Diffusion Models Meet Contextual Bandits with Large Action Spaces

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Efficient exploration in contextual bandits is crucial due to their large action space, where uninformed exploration can lead to computational and statistical inefficiencies. However, the rewards of actions are often correlated, which can be leveraged for more efficient exploration. In this work, we use pre-trained diffusion model priors to capture these correlations and develop diffusion Thompson sampling (`dTS`). We establish both theoretical and algorithmic foundations for `dTS`. Specifically, we derive efficient posterior approximations (required by `dTS`) under a diffusion model prior, which are of independent interest beyond bandits and reinforcement learning. We analyze `dTS` in linear instances and provide a Bayes regret bound highlighting the benefits of using diffusion models as priors. Our experiments validate our theory and demonstrate `dTS`'s favorable performance.

1 Introduction

A *contextual bandit* is a popular and practical framework for online learning under uncertainty [Li et al., 2010]. In each round, an agent observes a *context*, takes an *action*, and receives a *reward* based on the context and action. The goal is to maximize the expected cumulative reward over n rounds, striking a balance between exploiting actions with high estimated rewards from available data and exploring other actions to improve current estimates. This trade-off is often addressed using either *upper confidence bound (UCB)* [Auer et al., 2002] or *Thompson sampling (TS)* [Scott, 2010].

The action space in contextual bandits is often large, resulting in less-than-optimal performance with standard exploration strategies. Luckily, actions usually exhibit correlations, making efficient exploration possible as one action may inform the agent about other actions. In particular, Thompson sampling offers remarkable flexibility, allowing its integration with informative priors [Hong et al., 2022b] that capture these correlations. Inspired by the achievements of diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020], which effectively approximate complex distributions [Dhariwal and Nichol, 2021, Rombach et al., 2022], this work captures action correlations by employing diffusion models as priors in contextual Thompson sampling.

We illustrate the idea using video streaming. The objective is to optimize watch time for a user j by selecting a video i from a catalog of K videos. Users j and videos i are associated with context vectors x_j and unknown video parameters θ_i, respectively. User j's expected watch time for video i is linear as $x_j^\top \theta_i$. Then, a natural strategy is to independently learn video parameters θ_i using `LinTS` or `LinUCB` [Agrawal and Goyal, 2013a, Abbasi-Yadkori et al., 2011], but this proves statistically inefficient for larger K. Fortunately, the reward when recommending a movie can provide informative insights into other movies. To capture this, we leverage offline estimates of video parameters denoted by $\hat{\theta}_i$ and build a diffusion model on them. This diffusion model approximates the video parameter distribution, capturing their dependencies. This model enriches contextual Thompson sampling as a prior, effectively capturing complex video dependencies while ensuring computational efficiency.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

We introduce a framework for contextual bandits with diffusion model priors, upon which we develop diffusion Thompson sampling (dTS) that is both computationally and statistically efficient. dTS requires *fast updates of the posterior* and *fast sampling from the posterior*, both of which are achieved through our novel efficient posterior approximations. These approximations become exact when both the diffusion model and likelihood are linear. We establish a bound on dTS's Bayes regret for this specific case, highlighting the advantages of using diffusion models as priors. Our empirical evaluations validate our theory and demonstrate dTS's strong performance across various settings.

Diffusion models were applied in offline decision-making [Ajay et al., 2022, Janner et al., 2022, Wang et al., 2022], but their use in online learning was only recently explored by Hsieh et al. [2023], who focused on *multi-armed bandits without theoretical guarantees*. Our work extends Hsieh et al. [2023] in two ways. First, we apply the concept to the broader contextual bandit, which is more practical and realistic. Second, we demonstrate that with diffusion models parametrized by linear score functions and linear rewards, we can derive exact closed-form posteriors without approximations. These exact posteriors are valuable as they enable theoretical analysis (unlike Hsieh et al. [2023], who did not provide theoretical guarantees) and motivate efficient approximations for non-linear score functions in contextual bandits, addressing gaps in Hsieh et al. [2023]'s focus on multi-armed bandits.

A key contribution, beyond applying diffusion models in contextual bandits, is the efficient *computation* and *sampling* of the posterior distribution of a d-dimensional parameter $\theta \mid H_t$, with H_t representing the data, when using a diffusion model prior on θ. This is relevant not only to bandits and reinforcement learning but also to a broader range of applications [Chung et al., 2022]. To motivate our approximations, we start with exact closed-form solutions for cases where both the score functions of the diffusion model and the likelihood are linear. These solutions form the basis for our approximations for non-linear score functions, demonstrating both strong empirical performance and computational efficiency. Our approach avoids the computational burden of heavy approximate sampling algorithms required for each latent parameter. For a detailed comparison with existing studies, see Appendix A, where we discuss diffusion models in decision-making, structured bandits, approximate posteriors, and more.

2 Setting

The agent interacts with a *contextual bandit* over n rounds. In round $t \in [n]$, the agent observes a *context* $X_t \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a *context space*, it takes an *action* $A_t \in [K]$, and then receives a stochastic reward $Y_t \in \mathbb{R}$ that depends on both the context X_t and the taken action A_t. Each action $i \in [K]$ is associated with an *unknown action parameter* $\theta_{*,i} \in \mathbb{R}^d$, so that the reward received in round t is $Y_t \sim P(\cdot \mid X_t; \theta_{*,A_t})$, where $P(\cdot \mid x; \theta_{*,i})$ is the reward distribution of action i in context x. Throughout the paper, we assume that the reward distribution is parametrized as a generalized linear model (GLM) [McCullagh and Nelder, 1989]. That is, for any $x \in \mathcal{X}$, $P(\cdot \mid x; \theta_{*,i})$ is an exponential-family distribution with mean $g(x^\top \theta_{*,i})$, where g is the mean function. For example, we recover linear bandits when $P(\cdot \mid x; \theta_{*,i}) = \mathcal{N}(\cdot; x^\top \theta_{*,i}, \sigma^2)$ where $\sigma > 0$ is the observation noise. Similarly, we recover logistic bandits [Filippi et al., 2010] if we let $g(u) = (1 + \exp(-u))^{-1}$ and $P(\cdot \mid x; \theta_{*,i}) = \mathrm{Ber}(g(x^\top \theta_{*,i}))$, where $\mathrm{Ber}(p)$ be the Bernoulli distribution with mean p.

We consider the *Bayesian* bandit setting [Russo and Van Roy, 2014, Hong et al., 2022b], where the action parameters $\theta_{*,i}$ are assumed to be sampled from a *known* prior distribution. We proceed to define this prior distribution using a diffusion model. The correlations between the action parameters $\theta_{*,i}$ are captured through a diffusion model, where they share a set of L consecutive *unknown latent parameters* $\psi_{*,\ell} \in \mathbb{R}^d$ for $\ell \in [L]$. Precisely, the action parameter $\theta_{*,i}$ depends on the L-th latent parameter $\psi_{*,L}$ as $\theta_{*,i} \mid \psi_{*,1} \sim \mathcal{N}(f_1(\psi_{*,1}), \Sigma_1)$, where the *score function* $f_1 : \mathbb{R}^d \to \mathbb{R}^d$ is *known*. Also, the $\ell-1$-th latent parameter $\psi_{*,\ell-1}$ depends on the ℓ-th latent parameter $\psi_{*,\ell}$ as $\psi_{*,\ell-1} \mid \psi_{*,\ell} \sim \mathcal{N}(f_\ell(\psi_{*,\ell}), \Sigma_\ell)$, where the score function $f_\ell : \mathbb{R}^d \to \mathbb{R}^d$ is known. Finally, the L-th latent parameter $\psi_{*,L}$ is sampled as $\psi_{*,L} \sim \mathcal{N}(0, \Sigma_{L+1})$. We summarize this model in (1) and its graph in Fig. 1.

$$
\begin{aligned}
\psi_{*,L} &\sim \mathcal{N}(0, \Sigma_{L+1})\,, && \qquad (1)\\
\psi_{*,\ell-1} \mid \psi_{*,\ell} &\sim \mathcal{N}(f_\ell(\psi_{*,\ell}), \Sigma_\ell)\,, && \forall \ell \in [L]/\{1\}\,,\\
\theta_{*,i} \mid \psi_{*,1} &\sim \mathcal{N}(f_1(\psi_{*,1}), \Sigma_1)\,, && \forall i \in [K]\,,\\
Y_t \mid X_t, \theta_{*,A_t} &\sim P(\cdot \mid X_t; \theta_{*,A_t})\,, && \forall t \in [n]\,.
\end{aligned}
$$

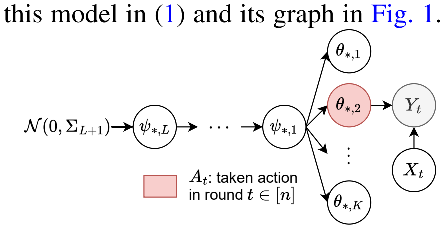

Figure 1: Graphical model of (1).

The model in (1) represents a Bayesian bandit, where the agent interacts with a bandit instance defined by $\theta_{*,i}$ over n rounds (4-th line in (1)). These action parameters $\theta_{*,i}$ are drawn from the generative process in the first 3 lines of (1). In practice, (1) can be built by pre-training a diffusion model on offline estimates of the action parameters $\theta_{*,i}$ [Hsieh et al., 2023].

A natural goal for the agent in this Bayesian framework is to minimize its *Bayes regret* [Russo and Van Roy, 2014] that measures the expected performance across multiple bandit instances $\theta_* = (\theta_{*,i})_{i \in [K]}$,

$$\mathcal{BR}(n) = \mathbb{E}\Big[\sum_{t=1}^{n} r(X_t, A_{t,*}; \theta_*) - r(X_t, A_t; \theta_*)\Big]\,, \tag{2}$$

where the expectation in (2) is taken over all random variables in (1). Here $r(x, i; \theta_*) = \mathbb{E}_{Y \sim P(\cdot|x;\theta_{*,i})}[Y]$ is the expected reward of action i in context x and $A_{t,*} = \arg\max_{i \in [K]} r(X_t, i; \theta_*)$ is the optimal action in round t. The Bayes regret is known to capture the benefits of using informative priors, and hence it is suitable for our problem.

3 Diffusion contextual Thompson sampling

We design Thompson sampling that samples the latent and action parameters hierarchically [Lindley and Smith, 1972]. Precisely, let $H_t = (X_k, A_k, Y_k)_{k \in [t-1]}$ be the history of all interactions up to round t and let $H_{t,i} = (X_k, A_k, Y_k)_{\{k \in [t-1]; A_k = i\}}$ be the history of interactions *with action* i up to round t. To motivate our algorithm, we decompose the posterior $\mathbb{P}(\theta_{*,i} = \theta \,|\, H_t)$ recursively as

$$\mathbb{P}(\theta_{*,i} = \theta \,|\, H_t) = \int_{\psi_{1:L}} Q_{t,L}(\psi_L) \prod_{\ell=2}^{L} Q_{t,\ell-1}(\psi_{\ell-1} \mid \psi_\ell) P_{t,i}(\theta \mid \psi_1)\, \mathrm{d}\psi_{1:L}\,, \quad \text{where} \tag{3}$$

$Q_{t,L}(\psi_L) = \mathbb{P}(\psi_{*,L} = \psi_L \,|\, H_t)$ is the *latent-posterior* density of $\psi_{*,L} \mid H_t$. Moreover, for any $\ell \in [2 : L]$, $Q_{t,\ell-1}(\psi_{\ell-1} \mid \psi_\ell) = \mathbb{P}(\psi_{*,\ell-1} = \psi_{\ell-1} \,|\, H_t, \psi_{*,\ell} = \psi_\ell)$ is the *conditional latent-posterior* density of $\psi_{*,\ell-1} \mid H_t, \psi_{*,\ell} = \psi_\ell$. Finally, for any action $i \in [K]$, $P_{t,i}(\theta \mid \psi_1) = \mathbb{P}(\theta_{*,i} = \theta \,|\, H_{t,i}, \psi_{*,1} = \psi_1)$ is the *conditional action-posterior* density of $\theta_{*,i} \mid H_{t,i}, \psi_{*,1} = \psi_1$.

The decomposition in (3) inspires hierarchical sampling. In round t, we initially sample the L-th latent parameter as $\psi_{t,L} \sim Q_{t,L}(\cdot)$. Then, for $\ell \in [L]/\{1\}$, we sample the $\ell-1$-th latent parameter given that $\psi_{*,\ell} = \psi_{t,\ell}$, as $\psi_{t,\ell-1} \sim Q_{t,\ell-1}(\cdot \mid \psi_{t,\ell})$. Lastly, given that $\psi_{*,1} = \psi_{t,1}$, each action parameter is sampled *individually* as $\theta_{t,i} \sim P_{t,i}(\theta \mid \psi_{t,1})$. This is possible because action parameters $\theta_{*,i}$ are conditionally independent given $\psi_{*,1}$. This leads to Algorithm 1, named **d**iffusion **T**hompson **S**ampling (`dTS`). `dTS` requires sampling from the $K + L$ posteriors $P_{t,i}$ and $Q_{t,\ell}$. Thus we start by providing an efficient recursive scheme to express these posteriors using known quantities. We note that these expressions do not necessarily lead to closed-form posteriors and approximation might be needed. First, the conditional action-posterior $P_{t,i}(\cdot \mid \psi_1)$ can be written as

$$P_{t,i}(\theta \mid \psi_1) \propto \prod_{k \in S_{t,i}} P(Y_k \mid X_k; \theta) \mathcal{N}(\theta; f_1(\psi_1), \Sigma_1)\,, \tag{4}$$

where $S_{t,i} = \{\ell \in [t-1], A_\ell = i\}$ are the rounds where the agent takes action i up to round t. Moreover, let $\mathcal{L}_\ell(\psi_\ell) = \mathbb{P}(H_t \,|\, \psi_{*,\ell} = \psi_\ell)$ be the likelihood of observations up to round t given that $\psi_{*,\ell} = \psi_\ell$. Then, for any $\ell \in [L]/\{1\}$, the $\ell-1$-th conditional latent-posterior $Q_{t,\ell-1}(\cdot \mid \psi_\ell)$ is

$$Q_{t,\ell-1}(\psi_{\ell-1} \mid \psi_\ell) \propto \mathcal{L}_{\ell-1}(\psi_{\ell-1}) \mathcal{N}(\psi_{\ell-1}, f_\ell(\psi_\ell), \Sigma_\ell)\,, \tag{5}$$

and $Q_{t,L}(\psi_L) \propto \mathcal{L}_L(\psi_L) \mathcal{N}(\psi_L, 0, \Sigma_{L+1})$. All the terms above are known, except the likelihoods $\mathcal{L}_\ell(\psi_\ell)$ for $\ell \in [L]$. These are computed recursively as follows. First, the basis of the recursion is

$$\mathcal{L}_1(\psi_1) = \prod_{i=1}^{K} \int_{\theta_i} \prod_{k \in S_{t,i}} P(Y_k \mid X_k; \theta_i) \mathcal{N}(\theta_i; f_1(\psi_1), \Sigma_1)\, \mathrm{d}\theta_i. \tag{6}$$

Then for $\ell \in [L]/\{1\}$, the recursive step is $\mathcal{L}_\ell(\psi_\ell) = \int_{\psi_{\ell-1}} \mathcal{L}_{\ell-1}(\psi_{\ell-1}) \mathcal{N}(\psi_{\ell-1}; f_\ell(\psi_\ell), \Sigma_\ell)\, \mathrm{d}\psi_{\ell-1}$.

All posterior expressions above use known quantities $(f_\ell, \Sigma_\ell, P(y \mid x; \theta))$. However, these expressions typically need to be approximated, except when the score functions f_ℓ are linear and the reward distribution $P(\cdot \mid x; \theta)$ is linear-Gaussian, where closed-form solutions can be obtained with careful derivations. These approximations are not trivial, and prior studies often rely on computationally intensive approximate sampling algorithms. In the following sections, we explain how we derive our efficient approximations which are motivated by the closed-form solutions of linear instances.

Algorithm 1 dTS: **d**iffusion **T**hompson **S**ampling

Input: Prior: $f_\ell, \ell \in [L]$, $\Sigma_\ell, \ell \in [L+1]$, and P.
for $t = 1, \dots, n$ **do**
  Sample $\psi_{t,L} \sim Q_{t,L}$ (requires fast approximate posterior update and sampling)
  for $\ell = L, \dots, 2$ **do**
    Sample $\psi_{t,\ell-1} \sim Q_{t,\ell-1}(\cdot \mid \psi_{t,\ell})$ (requires fast approximate posterior update and sampling)
  for $i = 1, \dots, K$ **do**
    Sample $\theta_{t,i} \sim P_{t,i}(\cdot \mid \psi_{t,1})$ (requires fast approximate posterior update and sampling)
  Take action $A_t = \text{argmax}_{i \in [K]} r(X_t, i; \theta_t)$, where $\theta_t = (\theta_{t,i})_{i \in [K]}$
  Receive reward $Y_t \sim P(\cdot \mid X_t; \theta_{*,A_t})$ and update posteriors $Q_{t+1,\ell}$ and $P_{t+1,i}$.

3.1 Linear diffusion model

Assume the score functions f_ℓ are linear such as $f_\ell(\psi_{*,\ell}) = \mathrm{W}_\ell \psi_{*,\ell}$ for $\ell \in [L]$, where $\mathrm{W}_\ell \in \mathbb{R}^{d \times d}$ are *known mixing matrices*. Then, (1) becomes a linear Gaussian system (LGS) [Bishop, 2006] in this case. This model is important, both in theory and practice. For theory, it leads to closed-form posteriors when the reward distribution is linear-Gaussian as $P(\cdot \mid x; \theta_{*,i}) = \mathcal{N}(\cdot; x^\top \theta_{*,i}, \sigma^2)$. This allows bounding the Bayes regret of dTS. For practice, the posterior expressions are used to motivate efficient approximations for the general case in (1) as we show in Section 3.2.

The reward distribution is parameterized as a generalized linear model (GLM) [McCullagh and Nelder, 1989], allowing for non-linear rewards. Thus, we need posterior approximation despite linearity in score functions. Since this non-linearity arises solely from the reward distribution, we approximate it by a Gaussian and propagate this approximation to the latent parameters. This results in efficient posterior approximations that are exact when the reward function is Gaussian (a special case of the GLM model). Specifically, the reward distribution $P(\cdot \mid x; \theta)$ is an exponential family distribution with a mean function denoted by g. Then, we approximate the corresponding likelihood as $\mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \approx \mathcal{N}\big(\theta; \hat{B}_{t,i}, \hat{G}_{t,i}^{-1}\big)$, where $\hat{B}_{t,i}$ and $\hat{G}_{t,i}$ are the maximum likelihood estimate (MLE) and the Hessian of the negative log-likelihood, respectively, and they are defined as

$$\hat{B}_{t,i} = \arg\max_{\theta \in \mathbb{R}^d} \log \mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right), \qquad \hat{G}_{t,i} = \textstyle\sum_{k \in S_{t,i}} \dot{g}\big(X_k^\top \hat{B}_{t,i}\big) X_k X_k^\top . \tag{7}$$

where $S_{t,i} = \{\ell \in [t-1] : A_\ell = i\}$ represents the rounds where the agent takes action i up to round t. This simple approximation makes all posteriors Gaussian. Specifically, the conditional action-posterior is Gaussian and is given by $P_{t,i}(\cdot \mid \psi_1) = \mathcal{N}(\cdot; \hat{\mu}_{t,i}, \hat{\Sigma}_{t,i})$, where $\hat{\mu}_{t,i}$ and $\hat{\Sigma}_{t,i}$ are computed using $\hat{B}_{t,i}$ and $\hat{G}_{t,i}$ in (7). Moreover, for $\ell \in [L-1]$, the ℓ-th conditional latent-posterior is also Gaussian, $Q_{t,\ell}(\cdot \mid \psi_{\ell+1}) = \mathcal{N}(\cdot; \bar{\mu}_{t,\ell}, \bar{\Sigma}_{t,\ell})$, where $\bar{\mu}_{t,\ell}$ and $\bar{\Sigma}_{t,\ell}$ are computed recursively. The recursion starts with $\bar{\mu}_{t,1}$ and $\bar{\Sigma}_{t,1}$, which are calculated using $\hat{B}_{t,i}$ and $\hat{G}_{t,i}$ in (7). Full expressions are provided in Appendix B.1. The only approximation made is $\mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \approx \mathcal{N}\big(\theta; \hat{B}_{t,i}, \hat{G}_{t,i}^{-1}\big)$, and we propagated it to latent posteriors. Thus, these posterior approximations become exact when the reward distribution follows a linear-Gaussian model, $P(\cdot \mid x; \theta_{*,a}) = \mathcal{N}(\cdot; x^\top \theta_{*,a}, \sigma^2)$.

3.2 Non-linear diffusion model

After deriving the posteriors for linear score functions, we return to the general model in (1). Approximation is needed since both the score functions and rewards can be non-linear. To avoid computational challenges, we use a simple and intuitive approximation, where all posteriors $P_{t,i}$ and $Q_{t,\ell}$ are approximated by Gaussians that are computed recursively. First, the conditional action-posterior is approximated by a Gaussian distribution as $P_{t,i}(\cdot \mid \psi_1) = \mathcal{N}(\cdot; \hat{\mu}_{t,i}, \hat{\Sigma}_{t,i})$, where

$$\hat{\Sigma}_{t,i}^{-1} = \Sigma_1^{-1} + \hat{G}_{t,i} \qquad\qquad \hat{\mu}_{t,i} = \hat{\Sigma}_{t,i}\big(\Sigma_1^{-1} f_1(\psi_1) + \hat{G}_{t,i} \hat{B}_{t,i}\big). \tag{8}$$

In the absence of samples, $G_{t,i} = 0_{d \times d}$. Thus, the approximate action posterior in (8) matches precisely the term $\mathcal{N}(f_1(\psi_1), \Sigma_1)$ in the diffusion prior (1). Moreover, as more data is accumulated, $G_{t,i}$ increases, and the influence of the prior diminishes as $\hat{G}_{t,i} \hat{B}_{t,i}$ will dominate the prior term $\Sigma_1^{-1} f_1(\psi_1)$. Similarly, for $\ell \in [L]/\{1\}$, the $\ell-1$-th conditional latent-posterior is approximated by a Gaussian distribution as $Q_{t,\ell-1}(\cdot \mid \psi_\ell) = \mathcal{N}(\bar{\mu}_{t,\ell-1}, \bar{\Sigma}_{t,\ell-1})$, where

$$\bar{\Sigma}_{t,\ell-1}^{-1} = \Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}\,, \qquad \bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\big(\Sigma_\ell^{-1} f_\ell(\psi_\ell) + \bar{B}_{t,\ell-1}\big)\,, \tag{9}$$

and the L-th latent-posterior is $Q_{t,L}(\cdot) = \mathcal{N}(\bar{\mu}_{t,L}, \bar{\Sigma}_{t,L})$,

$$\bar{\Sigma}_{t,L}^{-1} = \Sigma_{L+1}^{-1} + \bar{G}_{t,L}\,, \qquad \bar{\mu}_{t,L} = \bar{\Sigma}_{t,L}\bar{B}_{t,L}\,. \tag{10}$$

Here, $\bar{G}_{t,\ell}$ and $\bar{B}_{t,\ell}$ for $\ell \in [L]$ are computed recursively. The basis of the recursion are

$$\bar{G}_{t,1} = \textstyle\sum_{i=1}^K \big(\Sigma_1^{-1} - \Sigma_1^{-1}\hat{\Sigma}_{t,i}\Sigma_1^{-1}\big)\,, \qquad \bar{B}_{t,1} = \Sigma_1^{-1}\sum_{i=1}^K \hat{\Sigma}_{t,i}\hat{G}_{t,i}\hat{B}_{t,i}\,. \tag{11}$$

Then, the recursive step for $\ell \in [L]/\{1\}$ is,

$$\bar{G}_{t,\ell} = \Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\,, \qquad \bar{B}_{t,\ell} = \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\bar{B}_{t,\ell-1}\,. \tag{12}$$

Similarly, in the absence of samples, $Q_{t,\ell-1}$ in (9) precisely matches the term $\mathcal{N}(f_\ell(\psi_1), \Sigma_\ell)$ in the diffusion prior (1). As more data is accumulated, the influence of this prior diminishes. Therefore, this approximation retains a key attribute of exact posteriors: they match the prior when there is no data, and the prior's effect diminishes as data accumulates.

4 Analysis

We analyze dTS under the linear diffusion model in Section 3.1 with linear rewards $P(\cdot \mid x; \theta_{*,a}) = \mathcal{N}(\cdot; x^\top\theta_{*,a}, \sigma^2)$. This assumption leads to a structure with L layers of linear Gaussian relationships, allowing for theory inspired by linear bandits [Agrawal and Goyal, 2013a, Abbasi-Yadkori et al., 2011]. However, proofs are not the same, and technical challenges remain (explained in Appendix D).

Although our result holds for milder assumptions, we make some simplifications for clarity and interpretability. We assume that **(A1)** Contexts satisfy $\|X_t\|_2^2 = 1$ for any $t \in [n]$. **(A2)** Mixing matrices and covariances satisfy $\lambda_1(\mathrm{W}_\ell^\top \mathrm{W}_\ell) = 1$ for any $\ell \in [L]$ and $\Sigma_\ell = \sigma_\ell^2 I_d$ for any $\ell \in [L+1]$. Note that **(A1)** can be relaxed to any contexts X_t with bounded norms $\|X_t\|_2$. Also, **(A2)** can be relaxed to positive definite covariances Σ_ℓ and arbitrary mixing matrices W_ℓ. In this section, we write $\tilde{\mathcal{O}}$ for the big-O notation up to polylogarithmic factors. We start by stating our bound for dTS.

Theorem 4.1. *Let $\sigma_{\text{MAX}}^2 = \max_{\ell \in [L+1]} 1 + \frac{\sigma_\ell^2}{\sigma^2}$. For any $\delta \in (0,1)$, the Bayes regret of* dTS *under Section 3.1 with linear rewards,* **(A1)** *and* **(A2)** *is bounded as*

$$\mathcal{BR}(n) \le \sqrt{2n\big(\mathcal{R}^{\text{ACT}}(n) + \textstyle\sum_{\ell=1}^L \mathcal{R}_\ell^{\text{LAT}}\big)\log(1/\delta)\big)} + cn\delta\,, \text{ with } c > 0 \text{ is constant and,} \tag{13}$$

$$\mathcal{R}^{\text{ACT}}(n) = c_0 dK \log\big(1 + \tfrac{n\sigma_1^2}{d}\big),\ c_0 = \tfrac{\sigma_1^2}{\log(1+\sigma_1^2)}\,, \quad \mathcal{R}_\ell^{\text{LAT}} = c_\ell d \log\big(1 + \tfrac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\big), c_\ell = \tfrac{\sigma_{\ell+1}^2\sigma_{\text{MAX}}^{2\ell}}{\log(1+\sigma_{\ell+1}^2)},$$

(13) holds for any $\delta \in (0,1)$. In particular, the term $cn\delta$ is constant when $\delta = 1/n$. Then, the bound is $\tilde{\mathcal{O}}(\sqrt{n})$, and this dependence on the horizon n aligns with prior Bayes regret bounds. The bound comprises $L+1$ main terms, $\mathcal{R}^{\text{ACT}}(n)$ and $\mathcal{R}_\ell^{\text{LAT}}$ for $\ell \in [L]$. First, $\mathcal{R}^{\text{ACT}}(n)$ relates to action parameters learning, conforming to a standard form [Lu and Van Roy, 2019]. Similarly, $\mathcal{R}_\ell^{\text{LAT}}$ is associated with learning the ℓ-th latent parameter. Roughly speaking, our bound captures that our problem can be seen as $L+1$ sequential linear bandit instances stacked upon each other.

Technical contributions. dTS uses hierarchical sampling. Thus the marginal posterior distribution of $\theta_{*,i} \mid H_t$ is not explicitly defined. The first contribution is deriving $\theta_{*,i} \mid H_t$ using the total covariance decomposition combined with an induction proof, as our posteriors in Section 3.1 were derived recursively. Unlike standard analyses where the posterior distribution of $\theta_{*,i} \mid H_t$ is predetermined due to the absence of latent parameters, our method necessitates this recursive total covariance decomposition. Moreover, in standard proofs, we need to quantify the increase in posterior precision for the action taken A_t in each round $t \in [n]$. However, in dTS, our analysis extends beyond this. We not only quantify the posterior information gain for the taken action but also for every latent parameter, since they are also learned. To elaborate, we use the recursive formulas in Section 3.1 that connect the posterior covariance of each latent parameter $\psi_{*,\ell}$ with the covariance of the posterior action parameters $\theta_{*,i}$. This allows us to propagate the information gain associated with the action taken in round A_t to all latent parameters $\psi_{*,\ell}$, for $\ell \in [L]$ by induction. Finally, we carefully bound the resulting terms so that the constants reflect the parameters of the linear diffusion model. More technical details are provided in Appendix D.

To include more structure, we propose the *sparsity* assumption **(A3)** $\mathrm{W}_\ell = (\bar{\mathrm{W}}_\ell, 0_{d,d-d_\ell})$, where $\bar{\mathrm{W}}_\ell \in \mathbb{R}^{d\times d_\ell}$ for any $\ell \in [L]$. Note that **(A3)** is not an assumption when $d_\ell = d$ for any $\ell \in [L]$. Notably, **(A3)** incorporates a plausible structural characteristic that a diffusion model could capture.

Proposition 4.2 (Sparsity)**.** *Let $\sigma_{\text{MAX}}^2 = \max_{\ell\in[L+1]} 1 + \frac{\sigma_\ell^2}{\sigma^2}$. For any $\delta \in (0,1)$, the Bayes regret of* `dTS` *under Section 3.1 with linear rewards,* **(A1)**, **(A2)** *and* **(A3)** *is bounded as*

$$\mathcal{BR}(n) \leq \sqrt{2n\big(\mathcal{R}^{\text{ACT}}(n) + \textstyle\sum_{\ell=1}^{L} \tilde{\mathcal{R}}_\ell^{\text{LAT}}\big)\log(1/\delta)\Big)} + cn\delta\,, \text{ with } c>0 \text{ is constant,} \tag{14}$$

$$\mathcal{R}^{\text{ACT}}(n) = c_0 dK \log\left(1+\frac{n\sigma_1^2}{d}\right), c_0 = \frac{\sigma_1^2}{\log\left(1+\sigma_1^2\right)}\,, \quad \tilde{\mathcal{R}}_\ell^{\text{LAT}} = c_\ell d_\ell \log\left(1+\frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\right), c_\ell = \frac{\sigma_{\ell+1}^2\sigma_{\text{MAX}}^{2\ell}}{\log\left(1+\sigma_{\ell+1}^2\right)}.$$

From Proposition 4.2, our bounds scales as $\mathcal{BR}(n) = \tilde{\mathcal{O}}\Big(\sqrt{n(dK\sigma_1^2 + \sum_{\ell=1}^{L} d_\ell \sigma_{\ell+1}^2 \sigma_{\text{MAX}}^{2\ell})}\Big)$. The Bayes regret bound has a clear interpretation: if the true environment parameters are drawn from the prior, then the expected regret of an algorithm stays below that bound. Consequently, a less informative prior (such as high variance) leads to a more challenging problem and thus a higher bound. Then, smaller values of K, L, d or d_ℓ translate to fewer parameters to learn, leading to lower regret. The regret also decreases when the initial variances σ_ℓ^2 decrease. These dependencies are common in Bayesian analysis, and empirical results match them. The reader might question the dependence of our bound on both L and K. We will address this next.

Why the bound increases with K? This arises due to our conditional learning of $\theta_{*,i}$ given $\psi_{*,1}$. Rather than assuming deterministic linearity, $\theta_{*,i} = \mathrm{W}_1\psi_{*,1}$, we account for stochasticity by modeling $\theta_{*,i} \sim \mathcal{N}(\mathrm{W}_1\psi_{*,1}, \sigma_1^2 I_d)$. This makes `dTS` robust to misspecification scenarios where $\theta_{*,i}$ is not perfectly linear with respect to $\psi_{*,1}$, at the cost of additional learning of $\theta_{*,i} \mid \psi_{*,1}$. If we were to assume deterministic linearity ($\sigma_1 = 0$), our regret bound would scale with L only.

Why the bound increases with L? This is because increasing the number of layers L adds more initial uncertainty due to the additional covariance introduced by the extra layers. However, this does not imply that we should always use $L = 1$ (the minimum possible L). While a higher L complicates online learning and increases regret bound, it also enables the capture of a more complex prior distribution through offline pre-training of the diffusion model. Thus, a trade-off exists in practice. A smaller L results in faster computation and easier learning for `dTS`, but the learned prior might deviate from reality, potentially violating the "true prior assumption" used to derive the regret bound. On the other hand, a larger L allows for better modeling of complex action distributions, producing a prior that more accurately reflects reality and strengthens the validity of the bound.

4.1 Discussion

Computational benefits. Action correlations prompt an intuitive approach: marginalize all latent parameters and maintain a joint posterior of $(\theta_{*,i})_{i\in[K]} \mid H_t$. Unfortunately, this is computationally inefficient for large action spaces. To illustrate, suppose that all posteriors are multivariate Gaussians (Section 3.1). Then maintaining the joint posterior $(\theta_{*,i})_{i\in[K]} \mid H_t$ necessitates converting and storing its $dK \times dK$-dimensional covariance matrix. Then the time and space complexities are $\mathcal{O}(K^3d^3)$ and $\mathcal{O}(K^2d^2)$. In contrast, the time and space complexities of `dTS` are $\mathcal{O}\big((L+K)d^3\big)$ and $\mathcal{O}\big((L+K)d^2\big)$. This is because `dTS` requires converting and storing $L+K$ covariance matrices, each being $d\times d$-dimensional. The improvement is huge when $K \gg L$, which is common in practice. Certainly, a more straightforward way to enhance computational efficiency is to discard latent parameters and maintain K individual posteriors, each relating to an action parameter $\theta_{*,i} \in \mathbb{R}^d$ (`LinTS`). This improves time and space complexity to $\mathcal{O}\big(Kd^3\big)$ and $\mathcal{O}\big(Kd^2\big)$, respectively. However, `LinTS` maintains independent posteriors and fails to capture the correlations among actions; it only models $\theta_{*,i} \mid H_{t,i}$ rather than $\theta_{*,i} \mid H_t$ as done by `dTS`. Consequently, `LinTS` incurs higher regret due to the information loss caused by unused interactions of similar actions. Our regret bound and empirical results reflect this aspect.

Statistical benefits. We do not provide a matching lower bound. The only Bayesian lower bound that we know of is $\Omega(\log^2(n))$ for a much simpler K-armed bandit [Lai, 1987, Theorem 3]. All seminal works on Bayesian bandits do not match it and providing such lower bounds on Bayes regret is still relatively unexplored (even in standard settings) compared to the frequentist one. Therefore, we argue that our bound reflects the overall structure of the problem by comparing dTS to algorithms that only partially use the structure or do not use it at all as follows.

The linear diffusion model in Section 3.1 can be transformed into a Bayesian linear model (LinTS) by marginalizing out the latent parameters; in which case the prior on action parameters becomes $\theta_{*,i} \sim \mathcal{N}(0, \Sigma)$, with the $\theta_{*,i}$ being not necessarily independent, and Σ is the marginal initial covariance of action parameters and it writes $\Sigma = \sigma_1^2 I_d + \sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top$ with $\mathrm{B}_\ell = \prod_{k=1}^{\ell} \mathrm{W}_k$. Then, it is tempting to directly apply LinTS to solve our problem. This approach will induce higher regret because the additional uncertainty of the latent parameters is accounted for in Σ despite integrating them. This causes the *marginal* action uncertainty Σ to be much higher than the *conditional* action uncertainty $\sigma_1^2 I_d$ in (3.1), since we have $\Sigma = \sigma_1^2 I_d + \sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top \succcurlyeq \sigma_1^2 I_d$. This discrepancy leads to higher regret, especially when K is large. This is due to LinTS needing to learn K independent d-dimensional parameters, each with a considerably higher initial covariance Σ. This is also reflected by our regret bound. To simply comparisons, suppose that $\sigma \geq \max_{\ell \in [L+1]} \sigma_\ell$ so that $\sigma_{\text{MAX}}^2 \leq 2$. Then the regret bounds of dTS (where we bound $\sigma_{\text{MAX}}^{2\ell}$ by 2^ℓ) and LinTS read

$$\texttt{dTS} : \tilde{\mathcal{O}}\big(\sqrt{n(dK\sigma_1^2 + \textstyle\sum_{\ell=1}^{L} d_\ell \sigma_{\ell+1}^2 2^\ell)}\big)\,, \qquad \texttt{LinTS} : \tilde{\mathcal{O}}\big(\sqrt{ndK(\sigma_1^2 + \textstyle\sum_{\ell=1}^{L} \sigma_{\ell+1}^2)}\big)\,.$$

Then regret improvements are captured by the variances σ_ℓ and the sparsity dimensions d_ℓ, and we proceed to illustrate this through the following scenarios.

(I) Decreasing variances. Assume that $\sigma_\ell = 2^\ell$ for any $\ell \in [L+1]$. Then, the regrets become

$$\texttt{dTS} : \tilde{\mathcal{O}}\big(\sqrt{n(dK + \textstyle\sum_{\ell=1}^{L} d_\ell 4^\ell))}\big)\,, \qquad \texttt{LinTS} : \tilde{\mathcal{O}}\big(\sqrt{ndK2^L)}\big)$$

Now to see the order of gain, assume the problem is high-dimensional ($d \gg 1$), and set $L = \log_2(d)$ and $d_\ell = \lfloor \frac{d}{2^\ell} \rfloor$. Then the regret of dTS becomes $\tilde{\mathcal{O}}\big(\sqrt{nd(K+L)}\big)$, and hence the multiplicative factor 2^L in LinTS is removed and replaced with a smaller additive factor L.

(II) Constant variances. Assume that $\sigma_\ell = 1$ for any $\ell \in [L+1]$. Then, the regrets become

$$\texttt{dTS} : \tilde{\mathcal{O}}\big(\sqrt{n(dK + \textstyle\sum_{\ell=1}^{L} d_\ell 2^\ell))}\big)\,, \qquad \texttt{LinTS} : \tilde{\mathcal{O}}\big(\sqrt{ndKL)}\big)$$

Similarly, let $L = \log_2(d)$, and $d_\ell = \lfloor \frac{d}{2^\ell} \rfloor$. Then dTS's regret is $\tilde{\mathcal{O}}\big(\sqrt{nd(K+L)}\big)$. Thus the multiplicative factor L in LinTS is removed and replaced with the additive factor L. By comparing this to **(I)**, the gain with decreasing variances is greater than with constant ones. In general, diffusion models use decreasing variances [Ho et al., 2020] and hence we expect great gains in practice. All observed improvements in this section could become even more pronounced when employing non-linear diffusion models. In our current analysis, we used linear diffusion models, and yet we can already discern substantial differences. Moreover, under non-linear diffusion (1), the latent parameters cannot be analytically marginalized, making LinTS with exact marginalization inapplicable. Finally, Appendix D.7 provide an additional comparison and connection to hierarchies with two levels.

Large action space aspect. dTS's regret bound scales with $K\sigma_1^2$ instead of $K \sum_\ell \sigma_\ell^2$, particularly beneficial when σ_1 is small, as often seen in diffusion models. Our regret bound and experiments show that dTS outperforms LinTS more distinctly when the action space becomes larger. Prior studies [Foster et al., 2020, Xu and Zeevi, 2020, Zhu et al., 2022] proposed bandit algorithms that do not scale with K. However, our setting differs significantly from theirs, explaining our inherent dependency on K when $\sigma_1 > 0$. Precisely, they assume a reward function of $r(x, i; \theta_*) = \phi(x, i)^\top \theta_*$, with a shared $\theta_* \in \mathbb{R}^d$ and a known mapping ϕ. In contrast, we consider $r(x, i; \theta_*) = x^\top \theta_{*,i}$, with $\theta_* = (\theta_{*,i})_{i \in [K]} \in \mathbb{R}^{dK}$, requiring the learning of K separate d-dimensional action parameters. In their setting, with the availability of ϕ, the regret of dTS would similarly be independent of K. However, obtaining such a mapping ϕ can be challenging as it needs to encapsulate complex context-action dependencies. Notably, our setting reflects a common practical scenario, such as in recommendation systems where each product is often represented by its unique embedding.

5 Experiments

We evaluate dTS using synthetic data, to validate our theory and test dTS in large action spaces. We omit semi-synthetic data [Riquelme et al., 2018] as they often result in small action spaces. This

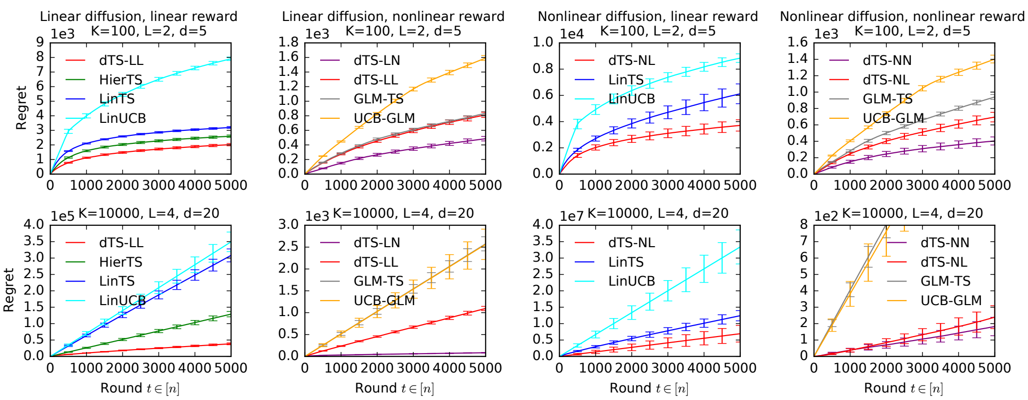

Figure 2: Regret of dTS with varying diffusion and reward models and varying parameters d, K, L.

choice is further justified by the fact that Hsieh et al. [2023] has already demonstrated the advantages of diffusion models in multi-armed bandits using such data, without theoretical guarantees.

5.1 Settings and baselines

We run 50 random simulations and plot the average regret with its standard error. We consider both linear and non-linear rewards. The distribution of linear rewards is $P(\cdot \mid x; \theta_a) = \mathcal{N}(x^\top \theta_a, \sigma^2)$ with $\sigma = 1$. The non-linear rewards are binary and generated from $P(\cdot \mid x; \theta_a) = \text{Ber}(g(x^\top \theta_a)))$, where g is the sigmoid function. The covariances are $\Sigma_\ell = I_d$, and the context X_t is uniformly drawn from $[-1, 1]^d$. We vary $d \in \{5, 20\}$, $L \in \{2, 4\}$ and $K \in \{10^2, 10^4\}$. We set the horizon $n = 5000$.

Linear diffusion. We consider the linear diffusion model in (3.1) where score functions are linear as $f_\ell(\psi) = \mathrm{W}_\ell \psi$ where W_ℓ are uniformly drawn from $[-1, 1]^{d \times d}$. To introduce sparsity, we zero out the last d_ℓ columns of W_ℓ, resulting in $\mathrm{W}_\ell = (\bar{\mathrm{W}}_\ell, 0_{d, d-d_\ell})$, where $(d_1, d_2) = (5, 2)$ when $d = 5$ and $L = 2$ and $(d_1, d_2, d_3, d_4) = (20, 10, 5, 2)$ when $d = 20$ and $L = 4$.

Non-linear diffusion. We consider the general diffusion model in (1) with score functions f_ℓ defined by two-layer neural networks with random weights in $[-1, 1]$, ReLU activation, and a hidden layer dimension of $h = 20$ when $d = 5$ and $h = 60$ when $d = 20$.

Baselines. When rewards are linear, we use LinUCB [Abbasi-Yadkori et al., 2011], LinTS [Agrawal and Goyal, 2013a], and HierTS [Hong et al., 2022b] that marginalizes out all latent parameters except $\psi_{*,L}$. This corresponds to HierTS-1 in Appendix D.7. When rewards are non-linear, we include UCB-GLM [Li et al., 2017], and GLM-TS [Chapelle and Li, 2012]. GLM-UCB [Filippi et al., 2010] induced high regret while HierTS was designed for linear rewards only and thus both are not included. We name dTS for each setting as dTS-dr, where the suffix d indicates the type of diffusion; L for linear and N for non-linear. The suffix r indicates the type of rewards; L for linear and N for non-linear. For instance, dTS-LL signifies dTS in linear diffusion (Section 3.1) with linear rewards.

5.2 Results and interpretations

Results are shown in Fig. 2 and we make the following observations:

1) dTS has better performance. dTS outperforms the baselines. First, when both the diffusion and rewards are linear, dTS-LL consistently outperforms all baselines that disregard the latent structure (LinTS and LinUCB) or incorporate it only partially (HierTS). Second, when the diffusion is linear and rewards are non-linear, dTS-LN surpasses all baselines. Third, when the diffusion is non-linear and rewards are linear, dTS-NL demonstrates significant performance gains compared to both LinTS and LinUCB. With non-linear diffusion and rewards, dTS-NN surpasses both GLM-TS and UCB-GLM.

2) Latent diffusion structure may be more important than the reward distribution. When rewards are non-linear (second and fourth columns in Fig. 2), we included variants of dTS that use the correct diffusion prior but the wrong reward distribution, employing linear-Gaussian instead of logistic-Bernoulli (dTS-LL in the second column and dTS-NL in the fourth column). In both cases, despite the misspecification of the reward distribution, these variants outperform models that use the correct reward distribution but neglect the latent diffusion structure, such as GLM-TS and UCB-GLM.

This underscores the significance of accounting for the latent structure, which can sometimes be more crucial than having an accurate reward distribution. Also, the performance gap between `dTS-NL` (non-linear diffusion) and `GLM-TS` and `UCB-GLM` is even more pronounced compared to the gap between `dTS-LL` (linear diffusion) and these baselines, possibly due to the increased complexity of the latent structure, in the non-linear diffusion, overshadowing the impact of the reward model itself.

3) Prior misspecification (Fig. 3). We consider a scenario where the prior used by `dTS` does not match the true prior. To simulate this, we use our setting with linear diffusion and rewards above, but the true parameters W_ℓ and Σ_ℓ are replaced by misspecified parameters $\mathrm{W}_\ell + \epsilon_1$ and $\Sigma_\ell + \epsilon_2$. Here, ϵ_1 and ϵ_2 are sampled uniformly from $[v, v{+}0.5]^{d\times d}$, with v controlling the level of misspecification. The higher the value of v, the greater the misspecification. We vary $v \in \{0.5, 1, 1.5\}$ and analyze its impact on `dTS`'s performance. For comparison, we include the well-specified `dTS-LL` and the most competitive baseline, `HierTS`. Results are shown in Fig. 3. As expected, `dTS`'s performance decreases with increasing misspecification. However, even with misspecification, `dTS` outperforms the most competitive baseline, except when $v = 1.5$, where their performances are comparable. Note that the entries of the true parameters W_ℓ and Σ_ℓ are smaller than 1, so values of $v \in \{0.5, 1, 1.5\}$ can lead to significant parameter misspecification. Yet, the performance of `dTS` with misspecified prior parameters remains favorable, suggesting that even an imperfect pre-trained diffusion model can be beneficial when used as prior.

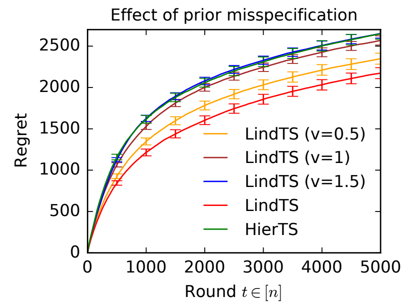

Figure 3: Prior misspecification effect.

4) Regret scaling with K, d and L matches our theory (Fig. 4). We verify the impact of the number of actions K, the context dimension d, and the diffusion depth L on the regret of `dTS`. We maintain the same experimental setup with linear diffusion and rewards, for which we have derived a Bayes regret upper bound. In Fig. 4, we plot the regret of `dTS-LL` across varying values of these parameters: $K \in \{10, 100, 500, 1000\}$, $d \in \{5, 10, 15, 20\}$, and $L \in \{2, 4, 5, 6\}$. As anticipated and aligned with our theory, the empirical regret increases as the values of K, d, or L grow. This trend arises because larger values of K, d, or L result in problem instances that are more challenging to learn, consequently leading to higher regret.

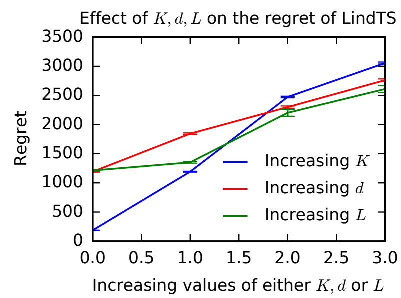

Figure 4: `dTS-LL`'s regret scaling.

5) Performance gap between `dTS` and `LinTS` widens as K increases (Fig. 5). To showcase `dTS`'s improved scalability to larger action spaces, we examine its performance across a range of K values, from 10 to 50,000, in our setting with linear diffusion and rewards. Fig. 5 reports the final cumulative regret for varying values of K for both `dTS-LL` and `LinTS`, observing that the gap in the performance becomes larger as K increases.

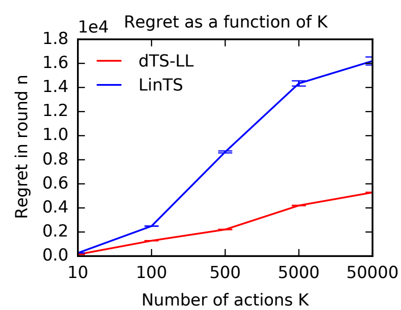

Figure 5: Regret of `dTS-LL` and `LinTS` with varying K.

6 Conclusion

Grappling with large action spaces in contextual bandits is challenging. Recognizing this, we focused on structured problems where action parameters are sampled from a diffusion model; upon which we built diffusion Thompson sampling (`dTS`). We developed both theoretical and algorithmic foundations for `dTS` in numerous practical settings. We identified several directions for future work. Exploring other approximations for non-linear diffusion models, both empirically and theoretically. From a theoretical perspective, future research could explore the advantages of non-linear diffusion models by deriving their Bayes regret bounds, akin to our analysis in Section 4. Empirically, investigating our and other approximations in complex tasks would be interesting. Additionally, exploring the extension of this work to offline (or off-policy) learning in contextual bandits [Swaminathan and Joachims, 2015, Aouali et al., 2023a] represents a promising avenue for future research.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013a.

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

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.

Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential smoothing for off-policy learning. In *International Conference on Machine Learning*, pages 984–1017. PMLR, 2023a.

Imad Aouali, Branislav Kveton, and Sumeet Katariya. Mixed-effect thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2087–2115. PMLR, 2023b.

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

Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In *Advances in Neural Information Processing Systems 26*, pages 2220–2228, 2013.

Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. Meta dynamic pricing: Transfer learning across experiments. *CoRR*, abs/1902.10918, 2019. URL https://arxiv.org/abs/1902.10918.

Soumya Basu, Branislav Kveton, Manzil Zaheer, and Csaba Szepesvari. No regrets for learning the prior in bandits. In *Advances in Neural Information Processing Systems 34*, 2021.

Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer, 2006.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Leonardo Cella, Karim Lounici, and Massimiliano Pontil. Multi-task representation learning with stochastic linear bandits. *arXiv preprint arXiv:2202.10066*, 2022.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2012.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Aniket Anand Deshmukh, Urun Dogan, and Clayton Scott. Multi-task learning for contextual bandits. In *Advances in Neural Information Processing Systems 30*, pages 4848–4856, 2017.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Sarah Filippi, Olivier Cappe, Aurelien Garivier, and Csaba Szepesvari. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594, 2010.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pages 757–765, 2014.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, pages 100–108, 2014.

Samarth Gupta, Shreyas Chaudhari, Subhojyoti Mukherjee, Gauri Joshi, and Osman Yagan. A unified approach to translate classical bandit algorithms to the structured bandit setting. *CoRR*, abs/1810.08164, 2018. URL https://arxiv.org/abs/1810.08164.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. In *Advances in Neural Information Processing Systems 33*, 2020.

Joey Hong, Branislav Kveton, Sumeet Katariya, Manzil Zaheer, and Mohammad Ghavamzadeh. Deep hierarchy in bandits. In *International Conference on Machine Learning*, pages 8833–8851. PMLR, 2022a.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical Bayesian bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022b.

Yu-Guan Hsieh, Shiva Prasad Kasiviswanathan, Branislav Kveton, and Patrick Blöbaum. Thompson sampling with diffusion generative prior. *arXiv preprint arXiv:2301.05182*, 2023.

Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pages 4349–4358. PMLR, 2021.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26, 2013.

John K Kruschke. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5): 658–676, 2010.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-Wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 1987.

Tor Lattimore and Remi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems 27*, pages 550–558, 2014.

Lihong Li, Wei Chu, John Langford, and Robert Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2071–2080, 2017.

Dennis Lindley and Adrian Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18, 1972.

Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, 2019.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pages 136–144, 2014.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

Amit Peleg, Naama Pearl, and Ron Meirr. Metalearning linear bandits by prior update. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Steven Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639 – 658, 2010.

Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel Hsu, Thodoris Lykouris, Miro Dudik, and Robert Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. In *Advances in Neural Information Processing Systems 34*, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.

Runzhe Wan, Lin Ge, and Rui Song. Metadata-based multi-task bandits with Bayesian hierarchical models. In *Advances in Neural Information Processing Systems 34*, 2021.

Runzhe Wan, Lin Ge, and Rui Song. Towards scalable and robust structured bandits: A meta-learning framework. *CoRR*, abs/2202.13227, 2022. URL https://arxiv.org/abs/2202.13227.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

Neil Weiss. *A Course in Probability*. Addison-Wesley, 2005.

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

Jiaqi Yang, Wei Hu, Jason D Lee, and Simon S Du. Impact of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*, 2020.

Tong Yu, Branislav Kveton, Zheng Wen, Ruiyi Zhang, and Ole Mengshoel. Graphical models meet bandits: A variational Thompson sampling approach. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pages 27428–27453. PMLR, 2022.

Supplementary materials

Notation. For any positive integer n, we define $[n] = \{1, 2, ..., n\}$. Let $v_1, \ldots, v_n \in \mathbb{R}^d$ be n vectors, $(v_i)_{i\in[n]} \in \mathbb{R}^{nd}$ is the nd-dimensional vector obtained by concatenating $v_1, \ldots, v_n$. For any matrix $\mathrm{A} \in \mathbb{R}^{d\times d}$, $\lambda_1(\mathrm{A})$ and $\lambda_d(\mathrm{A})$ denote the maximum and minimum eigenvalues of A, respectively. Finally, we write $\tilde{\mathcal{O}}$ for the big-O notation up to polylogarithmic factors.

A Extended related work

Thompson sampling (TS) operates within the Bayesian framework and it involves specifying a prior/likelihood model. In each round, the agent samples unknown model parameters from the current posterior distribution. The chosen action is the one that maximizes the resulting reward. TS is naturally randomized, particularly simple to implement, and has highly competitive empirical performance in both simulated and real-world problems [Russo and Van Roy, 2014, Chapelle and Li, 2012]. Regret guarantees for the TS heuristic remained open for decades even for simple models. Recently, however, significant progress has been made. For standard multi-armed bandits, TS is optimal in the Beta-Bernoulli model [Kaufmann et al., 2012, Agrawal and Goyal, 2013b], Gaussian-Gaussian model [Agrawal and Goyal, 2013b], and in the exponential family using Jeffrey's prior [Korda et al., 2013]. For linear bandits, TS is nearly-optimal [Russo and Van Roy, 2014, Agrawal and Goyal, 2017, Abeille and Lazaric, 2017]. In this work, we build TS upon complex diffusion priors and analyze the resulting Bayes regret [Russo and Van Roy, 2014] in the linear contextual bandit setting.

Decision-making with diffusion models gained attention recently, especially in offline learning [Ajay et al., 2022, Janner et al., 2022, Wang et al., 2022]. However, their application in online learning was only examined by Hsieh et al. [2023], which focused on meta-learning in multi-armed bandits without theoretical guarantees. In this work, we expand the scope of Hsieh et al. [2023] to encompass the broader contextual bandit framework. In particular, we provide theoretical analysis for linear instances, effectively capturing the advantages of using diffusion models as priors in contextual Thompson sampling. These linear cases are particularly captivating due to closed-form posteriors, enabling both theoretical analysis and computational efficiency; an important practical consideration.

Hierarchical Bayesian bandits [Bastani et al., 2019, Kveton et al., 2021, Basu et al., 2021, Simchowitz et al., 2021, Wan et al., 2021, Hong et al., 2022b, Peleg et al., 2022, Wan et al., 2022, Aouali et al., 2023b] applied TS to simple graphical models, wherein action parameters are generally sampled from a Gaussian distribution centered at a single latent parameter. These works mostly span meta- and multi-task learning for multi-armed bandits, except in cases such as Aouali et al. [2023b], Hong et al. [2022a] that consider the contextual bandit setting. Precisely, Aouali et al. [2023b] assume that action parameters are sampled from a Gaussian distribution centered at a linear mixture of multiple latent parameters. On the other hand, Hong et al. [2022a] applied TS to a graphical model represented by a tree. Our work can be seen as an extension of all these works to much more complex graphical models, for which both theoretical and algorithmic foundations are developed. Note that the settings in most of these works can be recovered with specific choices of the diffusion depth L and functions f_ℓ. This attests to the modeling power of dTS.

Approximate Thompson sampling is a major problem in the Bayesian inference literature. This is because most posterior distributions are intractable, and thus practitioners must resort to sophisticated computational techniques such as Markov chain Monte Carlo [Kruschke, 2010]. Prior works [Riquelme et al., 2018, Chapelle and Li, 2012, Kveton et al., 2020] highlight the favorable empirical performance of approximate Thompson sampling. Particularly, [Kveton et al., 2020] provide theoretical guarantees for Thompson sampling when using the Laplace approximation in generalized linear bandits (GLB). In our context, we incorporate approximate sampling when the reward exhibits non-linearity. While our approximation does not come with formal guarantees, it enjoys strong practical performance. An in-depth analysis of this approximation is left as a direction for future works. Similarly, approximating the posterior distribution when the diffusion model is non-linear as well as analyzing it is an interesting direction of future works.

Bandits with underlying structure also align with our work, where we assume a structured relationship among actions, captured by a diffusion model. In latent bandits [Maillard and Mannor, 2014, Hong et al., 2020], a single latent variable indexes multiple candidate models. Within structured finite-armed bandits [Lattimore and Munos, 2014, Gupta et al., 2018], each action is linked to a known mean function parameterized by a common latent parameter. This latent parameter is learned. TS was also applied to complex structures [Yu et al., 2020, Gopalan et al., 2014]. However, simultaneous computational and statistical efficiencies aren't guaranteed. Meta- and multi-task learning with upper confidence bound (UCB) approaches have a long history in bandits [Azar et al., 2013, Gentile et al., 2014, Deshmukh et al., 2017, Cella et al., 2020]. These, however, often adopt a frequentist perspective, analyze a stronger form of regret, and sometimes result in conservative algorithms. In contrast, our approach is Bayesian, with analysis centered on Bayes regret. Remarkably, our algorithm, `dTS`, performs well as analyzed without necessitating additional tuning. Finally, **Low-rank bandits** [Hu et al., 2021, Cella et al., 2022, Yang et al., 2020] also relate to our linear diffusion model when $L = 1$. Broadly, there exist two key distinctions between these prior works and the special case of our model (linear diffusion model with $L = 1$). First, they assume $\theta_{*,i} = \mathrm{W}_1 \psi_{*,1}$, whereas we incorporate additional uncertainty in the covariance Σ_1 to account for possible misspecification as $\theta_{*,i} = \mathcal{N}(\mathrm{W}_1 \psi_{*,1}, \Sigma_1)$. Consequently, these algorithms might suffer linear regret due to model misalignment. Second, we assume that the mixing matrix W_1 is available and pre-learned offline, whereas they learn it online. While this is more general, it leads to computationally expensive methods that are difficult to employ in a real-world online setting.

Large action spaces. Roughly speaking, the regret bound of `dTS` scales with $K\sigma_1^2$ rather than $K \sum_\ell \sigma_\ell^2$. This is particularly beneficial when σ_1 is small, a common scenario in diffusion models with decreasing variances. A notable case is when $\sigma_1 = 0$, where the regret becomes independent of K. Also, our analysis (Section 4.1) indicates that the gap in performance between `dTS` and `LinTS` becomes more pronounced when the number of action increases, highlighting `dTS`'s suitability for large action spaces. Note that some prior works [Foster et al., 2020, Xu and Zeevi, 2020, Zhu et al., 2022] proposed bandit algorithms that do not scale with K. However, our setting differs significantly from theirs, explaining our inherent dependency on K when $\sigma_1 > 0$. Precisely, they assume a reward function of $r(x, i) = \phi(x, i)^\top \theta_*$, with a shared $\theta_* \in \mathbb{R}^d$ across actions and a known mapping ϕ. In contrast, we consider $r(x, i) = x^\top \theta_{*,i}$, requiring the learning of K separate d-dimensional action parameters. In their setting, with the availability of ϕ, the regret of `dTS` would similarly be independent of K. However, obtaining such a mapping ϕ can be challenging as it needs to encapsulate complex context-action dependencies. Notably, our setting reflects a common practical scenario, such as in recommendation systems where each product is often represented by its embedding. In summary, the dependency on K is more related to our setting than the method itself, and `dTS` would scale with d only in their setting. Note that `dTS` is both computationally and statistically efficient (Section 4.1). This becomes particularly notable in large action spaces. Our empirical results in Fig. 2, notably with $K = 10^4$, demonstrate that `dTS` significantly outperforms the baselines. More importantly, the performance gap between `dTS` and these baselines is larger when the number of actions (K) increases, highlighting the improved scalability of `dTS` to large action spaces.

B Posterior derivations for linear diffusion models

Here, we assume the score functions f_ℓ are linear such as $f_\ell(\psi_{*,\ell}) = \mathrm{W}_\ell \psi_{*,\ell}$ for $\ell \in [L]$, where $\mathrm{W}_\ell \in \mathbb{R}^{d \times d}$ are *known mixing matrices*. Then, (1) becomes a linear Gaussian system (LGS) [Bishop, 2006] and can be summarized as follows

$$\begin{aligned}
\psi_{*,L} &\sim \mathcal{N}(0, \Sigma_{L+1})\,, && \\
\psi_{*,\ell-1} \mid \psi_{*,\ell} &\sim \mathcal{N}(\mathrm{W}_\ell \psi_{*,\ell}, \Sigma_\ell)\,, && \forall \ell \in [L]/\{1\}\,, \\
\theta_{*,i} \mid \psi_{*,1} &\sim \mathcal{N}(\mathrm{W}_1 \psi_{*,1}, \Sigma_1)\,, && \forall i \in [K]\,, \\
Y_t \mid X_t, \theta_{*,A_t} &\sim P(\cdot \mid X_t; \theta_{*,A_t})\,, && \forall t \in [n]\,.
\end{aligned} \tag{15}$$

In this section, we derive the $K + L$ posteriors $P_{t,i}$ and $Q_{t,\ell}$, for which we provide the full expressions in Appendix B.1. In our proofs, $p(x) \propto f(x)$ means that the probability density p satisfies $p(x) = \frac{f(x)}{Z}$ for any $x \in \mathbb{R}^d$, where Z is a normalization constant. In particular, we extensively use that if $p(x) \propto \exp[-\frac{1}{2} x^\top \Lambda x + x^\top m]$, where Λ is positive definite. Then p is the multivariate Gaussian density with covariance $\Sigma = \Lambda^{-1}$ and mean $\mu = \Sigma m$. These are standard notations and techniques to manipulate Gaussian distributions [Koller and Friedman, 2009, Chapter 7].

B.1 Posterior expressions for linear diffusion models

Recall that we posit that the reward distribution is parameterized as a generalized linear model (GLM) [McCullagh and Nelder, 1989], allowing for non-linear rewards. As a result, despite linearity in score functions, the non-linearity in rewards makes it challenging to obtain closed-form posteriors. However, since this non-linearity arises solely from the reward distribution, we approximate it using a Gaussian distribution. This leads to efficient posterior approximations that are exact in cases where the reward function is indeed Gaussian (a special case of the GLM model). Precisely, the reward distribution $P(\cdot \mid x; \theta)$ is an exponential-family distribution. Therefore, the log-likelihoods write $\log \mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) = \sum_{k \in S_{t,i}} Y_k X_k^\top \theta - A(X_k^\top \theta) + C(Y_k)$, where C is a real function, and A is a twice continuously differentiable function whose derivative is the mean function, $\dot{A} = g$. Now we let $\hat{B}_{t,i}$ and $\hat{G}_{t,i}$ be the maximum likelihood estimate (MLE) and the Hessian of the negative log-likelihood, respectively, defined as

$$\hat{B}_{t,i} = \arg\max_{\theta \in \mathbb{R}^d} \log \mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) , \qquad \hat{G}_{t,i} = \sum_{k \in S_{t,i}} \dot{g}\big(X_k^\top \hat{B}_{t,i}\big) X_k X_k^\top . \tag{16}$$

where $S_{t,i} = \{\ell \in [t-1] : A_\ell = i\}$ are the rounds where the agent takes action i up to round t. Then we approximation the respective likelihood as $\mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \approx \mathcal{N}\big(\theta; \hat{B}_{t,i}, \hat{G}_{t,i}^{-1}\big)$. This approximation makes all posteriors Gaussian. First, the conditional action-posterior reads $P_{t,i}(\cdot \mid \psi_1) = \mathcal{N}(\cdot; \hat{\mu}_{t,i}, \hat{\Sigma}_{t,i})$,

$$\hat{\Sigma}_{t,i}^{-1} = \Sigma_1^{-1} + \hat{G}_{t,i} \qquad \hat{\mu}_{t,i} = \hat{\Sigma}_{t,i}\big(\Sigma_1^{-1} \mathrm{W}_1 \psi_1 + \hat{G}_{t,i} \hat{B}_{t,i}\big). \tag{17}$$

For $\ell \in [L]/\{1\}$, the $\ell - 1$-th conditional latent-posterior is $Q_{t,\ell-1}(\cdot \mid \psi_\ell) = \mathcal{N}(\bar{\mu}_{t,\ell-1}, \bar{\Sigma}_{t,\ell-1})$,

$$\bar{\Sigma}_{t,\ell-1}^{-1} = \Sigma_\ell^{-1} + \bar{G}_{t,\ell-1} , \qquad \bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\big(\Sigma_\ell^{-1} \mathrm{W}_\ell \psi_\ell + \bar{B}_{t,\ell-1}\big) , \tag{18}$$

and the L-th latent-posterior is $Q_{t,L}(\cdot) = \mathcal{N}(\bar{\mu}_{t,L}, \bar{\Sigma}_{t,L})$,

$$\bar{\Sigma}_{t,L}^{-1} = \Sigma_{L+1}^{-1} + \bar{G}_{t,L} , \qquad \bar{\mu}_{t,L} = \bar{\Sigma}_{t,L} \bar{B}_{t,L} . \tag{19}$$

Finally, $\bar{G}_{t,\ell}$ and $\bar{B}_{t,\ell}$ for $\ell \in [L]$ are computed recursively. The basis of the recursion are

$$\bar{G}_{t,1} = \mathrm{W}_1^\top \sum_{i=1}^{K} \big(\Sigma_1^{-1} - \Sigma_1^{-1} \hat{\Sigma}_{t,i} \Sigma_1^{-1}\big) \mathrm{W}_1 , \qquad \bar{B}_{t,1} = \mathrm{W}_1^\top \Sigma_1^{-1} \sum_{i=1}^{K} \hat{\Sigma}_{t,i} \hat{G}_{t,i} \hat{B}_{t,i} . \tag{20}$$

Then, the recursive step for $\ell \in [L]/\{1\}$ is,

$$\bar{G}_{t,\ell} = \mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1} \bar{\Sigma}_{t,\ell-1} \Sigma_\ell^{-1}\big) \mathrm{W}_\ell , \qquad \bar{B}_{t,\ell} = \mathrm{W}_\ell^\top \Sigma_\ell^{-1} \bar{\Sigma}_{t,\ell-1} \bar{B}_{t,\ell-1} . \tag{21}$$

This concludes the derivation of our posterior approximation. Note that these approximations are exact when the reward distribution follows a linear-Gaussian model, $P(\cdot \mid x; \theta_{*,a}) = \mathcal{N}(\cdot; x^\top \theta_{*,a}, \sigma^2)$.

B.2 Derivation of Action-Posteriors for Linear Diffusion Models

To simplify derivations, we consider the case where the reward distribution is indeed linear-Gaussian as $P(\cdot \mid X_t; \theta_{*,A_t}) = \mathcal{N}\left(X_t^\top \theta_{*,A_t}, \sigma^2\right)$, but the same derivations can be applied when the rewards are non-linear. In this case, the likelihood approximation in (16) becomes exact as we have that $\mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \propto \mathcal{N}\big(\theta; \hat{B}_{t,i}, \hat{G}_{t,i}^{-1}\big)$, where $\hat{B}_{t,i}$ is the corresponding MLE and $\hat{G}_{t,i} = \sigma^{-2} \sum_{k \in S_{t,i}} X_k X_k^\top$ in this case. Our derivations rely on the fact that the MLE $\hat{B}_{t,i}$ in this linear-Gaussian case satisfies: $\hat{G}_{t,i} \hat{B}_{t,i} = v \sum_{k \in S_{t,i}} X_k Y_k^\top$.

Proposition B.1. *Consider the following model, which corresponds to the last two layers in Eq. (15)*

$$\begin{aligned} \theta_{*,i} \mid \psi_{*,1} &\sim \mathcal{N}\left(\mathrm{W}_1 \psi_{*,1}, \Sigma_1\right) , \\ Y_t \mid X_t, \theta_{*,A_t} &\sim \mathcal{N}\left(X_t^\top \theta_{*,A_t}, \sigma^2\right) , \qquad \forall t \in [n] . \end{aligned}$$

Then we have that for any $t \in [n]$ and $i \in [K]$, $P_{t,i}(\theta \mid \psi_1) = \mathbb{P}\left(\theta_{*,i} = \theta \,|\, \psi_{*,1} = \psi_1, H_{t,i}\right) = \mathcal{N}(\theta; \hat{\mu}_{t,i}, \hat{\Sigma}_{t,i})$, where

$$\hat{\Sigma}_{t,i}^{-1} = \hat{G}_{t,i} + \Sigma_1^{-1} , \qquad \hat{\mu}_{t,i} = \hat{\Sigma}_{t,i}\left(\hat{G}_{t,i} \hat{B}_{t,i} + \Sigma_1^{-1} \mathrm{W}_1 \psi_1\right) .$$

Proof. Let $v = \sigma^{-2}\,, \quad \Lambda_1 = \Sigma_1^{-1}$. Then the action-posterior decomposes as

$$\begin{aligned}
P_{t,i}(\theta \mid \psi_1) &= \mathbb{P}\left(\theta_{*,i} = \theta \,|\, \psi_{*,1} = \psi_1, H_{t,i}\right)\,, \\
&\propto \mathbb{P}\left(H_{t,i} \,|\, \psi_{*,1} = \psi_1, \theta_{*,i} = \theta\right) \mathbb{P}\left(\theta_{*,i} = \theta \,|\, \psi_{*,1} = \psi_1\right)\,, \quad \text{(Bayes rule)} \\
&= \mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \mathbb{P}\left(\theta_{*,i} = \theta \,|\, \psi_{*,1} = \psi_1\right)\,, \text{(given } \theta_{*,i}, H_{t,i} \text{ is independent of } \psi_{*,1}) \\
&= \prod_{k \in S_{t,i}} \mathcal{N}(Y_k; X_k^\top \theta, \sigma^2) \mathcal{N}(\theta; \mathrm{W}_1 \psi_1, \Sigma_1)\,, \\
&= \exp\Big[-\frac{1}{2}\Big(v \sum_{k \in S_{t,i}} (Y_k^2 - 2 Y_k X_k^\top \theta + (X_k^\top \theta)^2) + \theta^\top \Lambda_1 \theta - 2\theta^\top \Lambda_1 \mathrm{W}_1 \psi_1 \\
&\qquad + \big(\mathrm{W}_1 \psi_1\big)^\top \Lambda_1 \big(\mathrm{W}_1 \psi_1\big)\Big)\Big]\,, \\
&\propto \exp\Big[-\frac{1}{2}\Big(\theta^\top (v \sum_{k \in S_{t,i}} X_k X_k^\top + \Lambda_1)\theta - 2\theta^\top \Big(v \sum_{k \in S_{t,i}} X_k Y_k + \Lambda_1 \mathrm{W}_1 \psi_1\Big)\Big)\Big]\,, \\
&\propto \mathcal{N}\big(\theta; \hat{\mu}_{t,i}, \hat{\Lambda}_{t,i}^{-1}\big)\,,
\end{aligned}$$

with $\hat{\Lambda}_{t,i} = v \sum_{k \in S_{t,i}} X_k X_k^\top + \Lambda_1\,,\ \hat{\Lambda}_{t,i}\hat{\mu}_{t,i} = v \sum_{k \in S_{t,i}} X_k Y_k + \Lambda_1 \mathrm{W}_1 \psi_1$. Using that, in this linear-Gaussian case, $\hat{G}_{t,i} = v \sum_{k \in S_{t,i}} X_k X_k^\top$ and $\hat{G}_{t,i}\hat{B}_{t,i} = v \sum_{k \in S_{t,i}} X_k Y_k$ concludes the proof. ☐

The same proof applies when the reward distribution is not linear-Gaussian, with the approximation $\mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \approx \mathcal{N}\big(\theta; \hat{B}_{t,i}, \hat{G}_{t,i}^{-1}\big)$. Using this approximation in the derivations above leads to the same results.

B.3 Derivation of recursive latent-posteriors for linear diffusion models

Again, to simplify derivations, we consider the case where the reward distribution is indeed linear-Gaussian as $P(\cdot \mid X_t; \theta_{*,A_t}) = \mathcal{N}\left(X_t^\top \theta_{*,A_t}, \sigma^2\right)$, but the same derivations can be applied when the rewards are non-linear.

Proposition B.2. *For any $\ell \in [L]/\{1\}$, the $\ell-1$-th conditional latent-posterior reads $Q_{t,\ell-1}(\cdot \mid \psi_\ell) = \mathcal{N}(\bar{\mu}_{t,\ell-1}, \bar{\Sigma}_{t,\ell-1})$, with*

$$\bar{\Sigma}_{t,\ell-1}^{-1} = \Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}\,, \qquad \bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\big(\Sigma_\ell^{-1} \mathrm{W}_\ell \psi_\ell + \bar{B}_{t,\ell-1}\big)\,, \tag{22}$$

and the L-th latent-posterior reads $Q_{t,L}(\cdot) = \mathcal{N}(\bar{\mu}_{t,L}, \bar{\Sigma}_{t,L})$, with

$$\bar{\Sigma}_{t,L}^{-1} = \Sigma_{L+1}^{-1} + \bar{G}_{t,L}\,, \qquad \bar{\mu}_{t,L} = \bar{\Sigma}_{t,L} \bar{B}_{t,L}\,. \tag{23}$$

Proof. Let $\ell \in [L]/\{1\}$. Then, Bayes rule yields that

$$Q_{t,\ell-1}(\psi_{\ell-1} \mid \psi_\ell) \propto \mathbb{P}\left(H_t \,|\, \psi_{*,\ell-1} = \psi_{\ell-1}\right) \mathcal{N}(\psi_{\ell-1}, \mathrm{W}_\ell \psi_\ell, \Sigma_\ell)\,,$$

But from Lemma B.3, we know that

$$\mathbb{P}\left(H_t \,|\, \psi_{*,\ell-1} = \psi_{\ell-1}\right) \propto \exp\Big[-\frac{1}{2}\psi_{\ell-1}^\top \bar{G}_{t,\ell-1} \psi_{\ell-1} + \psi_{\ell-1}^\top \bar{B}_{t,\ell-1}\Big]\,.$$

Therefore,

$$\begin{aligned}
Q_{t,\ell-1}(\psi_{\ell-1} \mid \psi_\ell) &\propto \exp\Big[-\frac{1}{2}\psi_{\ell-1}^\top \bar{G}_{t,\ell-1} \psi_{\ell-1} + \psi_{\ell-1}^\top \bar{B}_{t,\ell-1}\Big] \mathcal{N}(\psi_{\ell-1}, \mathrm{W}_\ell \psi_\ell, \Sigma_\ell)\,, \\
&\propto \exp\Big[-\frac{1}{2}\psi_{\ell-1}^\top \bar{G}_{t,\ell-1} \psi_{\ell-1} + \psi_{\ell-1}^\top \bar{B}_{t,\ell-1} \\
&\qquad - \frac{1}{2}(\psi_{\ell-1} - \mathrm{W}_\ell \psi_\ell)^\top \Sigma_\ell^{-1} (\psi_{\ell-1} - \mathrm{W}_\ell \psi_\ell)\Big]\,, \\
&\overset{(i)}{\propto} \exp\Big[-\frac{1}{2}\psi_{\ell-1}^\top (\bar{G}_{t,\ell-1} + \Sigma_\ell^{-1}) \psi_{\ell-1} + \psi_{\ell-1}^\top (\bar{B}_{t,\ell-1} + \Sigma_\ell^{-1} \mathrm{W}_\ell \psi_\ell)\Big]\,, \\
&\overset{(ii)}{\propto} \mathcal{N}(\psi_{\ell-1}; \bar{\mu}_{t,\ell-1}, \bar{\Sigma}_{t,\ell-1})\,,
\end{aligned}$$

with $\bar{\Sigma}_{t,\ell-1}^{-1} = \Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}$ and $\bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\big(\Sigma_\ell^{-1}\mathrm{W}_\ell\psi_\ell + \bar{B}_{t,\ell-1}\big)$. In (i), we omit terms that are constant in $\psi_{\ell-1}$. In (ii), we complete the square. This concludes the proof for $\ell \in [L]/\{1\}$. For $Q_{t,L}$, we use Bayes rule to get

$$Q_{t,L}(\psi_L) \propto \mathbb{P}\left(H_t \,|\, \psi_{*,L} = \psi_L\right) \mathcal{N}(\psi_L, 0, \Sigma_{L+1})\,.$$

Then from Lemma B.3, we know that

$$\mathbb{P}\left(H_t \,|\, \psi_{*,L} = \psi_L\right) \propto \exp\Big[-\frac{1}{2}\psi_L^\top \bar{G}_{t,L}\psi_L + \psi_L^\top \bar{B}_{t,L}\Big]\,,$$

We then use the same derivations above to compute the product $\exp\Big[-\frac{1}{2}\psi_L^\top \bar{G}_{t,L}\psi_L + \psi_L^\top \bar{B}_{t,L}\Big] \times \mathcal{N}(\psi_L, 0, \Sigma_{L+1})$, which concludes the proof. □

Lemma B.3. *The following holds for any $t \in [n]$ and $\ell \in [L]$,*

$$\mathbb{P}\left(H_t \,|\, \psi_{*,\ell} = \psi_\ell\right) \propto \exp\Big[-\frac{1}{2}\psi_\ell^\top \bar{G}_{t,\ell}\psi_\ell + \psi_\ell^\top \bar{B}_{t,\ell}\Big]\,,$$

where $\bar{G}_{t,\ell}$ and $\bar{B}_{t,\ell}$ are defined by recursion in Section 3.1.

Proof. We prove this result by induction. To reduce clutter, we let $v = \sigma^{-2}$, and $\Lambda_1 = \Sigma_1^{-1}$. We start with the base case of the induction when $\ell = 1$.

(I) Base case. Here we want to show that $\mathbb{P}\left(H_t \,|\, \psi_{*,1} = \psi_1\right) \propto \exp\Big[-\frac{1}{2}\psi_1^\top \bar{G}_{t,1}\psi_1 + \psi_1^\top \bar{B}_{t,1})\Big]$, where $\bar{G}_{t,1}$ and $\bar{B}_{t,1}$ are given in Eq. (20). First, we have that

$$\begin{aligned}
\mathbb{P}\left(H_t \,|\, \psi_{*,1} = \psi_1\right) &\overset{(i)}{=} \prod_{i\in[K]} \mathbb{P}\left(H_{t,i} \,|\, \psi_{*,1} = \psi_1\right) = \prod_{i\in[K]} \int_\theta \mathbb{P}\left(H_{t,i}, \theta_{*,i} = \theta \,|\, \psi_{*,1} = \psi_1\right) \mathrm{d}\theta\,,\\
&= \prod_{i\in[K]} \int_\theta \mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \mathcal{N}\left(\theta; \mathrm{W}_1\psi_1, \Sigma_1\right) \mathrm{d}\theta\,,\\
&= \prod_{i\in[K]} \underbrace{\int_\theta \Big(\prod_{k\in S_{t,i}} \mathcal{N}(Y_k; X_k^\top\theta, \sigma^2)\Big)\mathcal{N}\left(\theta; \mathrm{W}_1\psi_1, \Sigma_1\right) \mathrm{d}\theta}_{h_i(\psi_1)}\,,\\
&= \prod_{i\in[K]} h_i(\psi_1)\,, \qquad (24)
\end{aligned}$$

where (i) follows from the fact that $\theta_{*,i}$ for $i \in [K]$ are conditionally independent given $\psi_{*,1} = \psi_1$ and that given $\theta_{*,i}$, $H_{t,i}$ is independent of $\psi_{*,1}$. Now we compute $h_i(\psi_1) = \int_\theta \Big(\prod_{k\in S_{t,i}} \mathcal{N}(Y_k; X_k^\top\theta, \sigma^2)\Big)\mathcal{N}\left(\theta; \mathrm{W}_1\psi_1, \Sigma_1\right) \mathrm{d}\theta$ as

$$\begin{aligned}
h_i(\psi_1) &= \int_\theta \Big(\prod_{k\in S_{t,i}} \mathcal{N}(Y_k; X_k^\top\theta, \sigma^2)\Big)\mathcal{N}(\theta; \mathrm{W}_1\psi_1, \Sigma_1)\,\mathrm{d}\theta\,,\\
&\propto \int_\theta \exp\Big[-\frac{1}{2}v\sum_{k\in S_{t,i}}(Y_k - X_k^\top\theta)^2 - \frac{1}{2}(\theta - \mathrm{W}_1\psi_1)^\top\Lambda_1(\theta - \mathrm{W}_1\psi_1)\Big]\,\mathrm{d}\theta\,,\\
&= \int_\theta \exp\Big[-\frac{1}{2}\Big(v\sum_{k\in S_{t,i}}(Y_k^2 - 2Y_k\theta^\top X_k + (\theta^\top X_k)^2) + \theta^\top\Lambda_1\theta - 2\theta^\top\Lambda_1\mathrm{W}_1\psi_1\\
&\qquad + (\mathrm{W}_1\psi_1)^\top\Lambda_1(\mathrm{W}_1\psi_1)\Big)\Big]\,\mathrm{d}\theta\,,\\
&\propto \int_\theta \exp\Big[-\frac{1}{2}\Big(\theta^\top\Big(v\sum_{k\in S_{t,i}}X_kX_k^\top + \Lambda_1\Big)\theta - 2\theta^\top\Big(v\sum_{k\in S_{t,i}}Y_kX_k\\
&\qquad + \Lambda_1\mathrm{W}_1\psi_1\Big) + (\mathrm{W}_1\psi_1)^\top\Lambda_1(\mathrm{W}_1\psi_1)\Big)\Big]\,\mathrm{d}\theta\,.
\end{aligned}$$

But we know that $\hat{G}_{t,i} = v \sum_{k \in S_{t,i}} X_k X_k^\top$, and $\hat{G}_{t,i}\hat{B}_{t,i} = v \sum_{k \in S_{t,i}} Y_k X_k$ (because we assumed linear-Gaussian likelihood). To further simplify expressions, we also let

$$V = \left(\hat{G}_{t,i} + \Lambda_1\right)^{-1}, \quad U = V^{-1}, \quad \beta = V\left(\hat{G}_{t,i}\hat{B}_{t,i} + \Lambda_1 \mathrm{W}_1 \psi_1\right).$$

We have that $UV = VU = I_d$, and thus

$$\begin{aligned}
h_i(\psi_1) &\propto \int_\theta \exp\left[-\frac{1}{2}\left(\theta^\top U \theta - 2\theta^\top UV\left(\hat{G}_{t,i}\hat{B}_{t,i} + \Lambda_1 \mathrm{W}_1\psi_1\right) + (\mathrm{W}_1\psi_1)^\top \Lambda_1 (\mathrm{W}_1\psi_1)\right)\right] \mathrm{d}\theta\,, \\
&= \int_\theta \exp\left[-\frac{1}{2}\left(\theta^\top U\theta - 2\theta^\top U\beta + (\mathrm{W}_1\psi_1)^\top\Lambda_1(\mathrm{W}_1\psi_1)\right)\right] \mathrm{d}\theta\,, \\
&= \int_\theta \exp\left[-\frac{1}{2}\left((\theta-\beta)^\top U(\theta-\beta) - \beta^\top U\beta + (\mathrm{W}_1\psi_1)^\top\Lambda_1(\mathrm{W}_1\psi_1)\right)\right] \mathrm{d}\theta\,, \\
&\propto \exp\left[-\frac{1}{2}\left(-\beta^\top U\beta + (\mathrm{W}_1\psi_1)^\top\Lambda_1(\mathrm{W}_1\psi_1)\right)\right]\,, \\
&= \exp\left[-\frac{1}{2}\left(-\left(\hat{G}_{t,i}\hat{B}_{t,i} + \Lambda_1\mathrm{W}_1\psi_1\right)^\top V\left(\hat{G}_{t,i}\hat{B}_{t,i} + \Lambda_1\mathrm{W}_1\psi_1\right) + (\mathrm{W}_1\psi_1)^\top\Lambda_1(\mathrm{W}_1\psi_1)\right)\right]\,, \\
&\propto \exp\left[-\frac{1}{2}\left(\psi_1^\top \mathrm{W}_1^\top\left(\Lambda_1 - \Lambda_1 V\Lambda_1\right)\mathrm{W}_1\psi_1 - 2\psi_1^\top\left(\mathrm{W}_1^\top\Lambda_1 V\hat{G}_{t,i}\hat{B}_{t,i}\right)\right)\right]\,, \\
&= \exp\left[-\frac{1}{2}\psi_1^\top\Omega_i\psi_1 + \psi_1^\top m_i\right]\,,
\end{aligned}$$

where

$$\begin{aligned}
\Omega_i &= \mathrm{W}_1^\top\left(\Lambda_1 - \Lambda_1 V\Lambda_1\right)\mathrm{W}_1 = \mathrm{W}_1^\top\left(\Lambda_1 - \Lambda_1(\hat{G}_{t,i}+\Lambda_1)^{-1}\Lambda_1\right)\mathrm{W}_1\,, \\
m_i &= \mathrm{W}_1^\top\Lambda_1 V\hat{G}_{t,i}\hat{B}_{t,i} = \mathrm{W}_1^\top\Lambda_1(\hat{G}_{t,i}+\Lambda_1)^{-1}\hat{G}_{t,i}\hat{B}_{t,i}\,.
\end{aligned} \tag{25}$$

But notice that $V = (\hat{G}_{t,i}+\Lambda_1)^{-1} = \hat{\Sigma}_{t,i}$ and thus

$$\Omega_i = \mathrm{W}_1^\top\left(\Lambda_1 - \Lambda_1\hat{\Sigma}_{t,i}\Lambda_1\right)\mathrm{W}_1\,, \qquad m_i = \mathrm{W}_1^\top\Lambda_1\hat{\Sigma}_{t,i}\hat{G}_{t,i}\hat{B}_{t,i}\,. \tag{26}$$

Finally, we plug this result in Eq. (24) to get

$$\begin{aligned}
\mathbb{P}\left(H_t \,|\, \psi_{*,1} = \psi_1\right) = \prod_{i\in[K]} h_i(\psi_1) &\propto \prod_{i\in[K]} \exp\left[-\frac{1}{2}\psi_1^\top\Omega_i\psi_1 + \psi_1^\top m_i\right]\,, \\
&= \exp\left[-\frac{1}{2}\psi_1^\top\sum_{i\in[K]}\Omega_i\psi_1 + \psi_1^\top\sum_{i\in[K]} m_i\right]\,, \\
&= \exp\left[-\frac{1}{2}\psi_1^\top\bar{G}_{t,1}\psi_1 + \psi_1^\top\bar{B}_{t,1}\right]\,,
\end{aligned}$$

where

$$\begin{aligned}
\bar{G}_{t,1} &= \sum_{i=1}^K \Omega_i = \sum_{i=1}^K \mathrm{W}_1^\top\left(\Lambda_1 - \Lambda_1\hat{\Sigma}_{t,i}\Lambda_1\right)\mathrm{W}_1 = \mathrm{W}_1^\top\sum_{i=1}^K\left(\Sigma_1^{-1} - \Sigma_1^{-1}\hat{\Sigma}_{t,i}\Sigma_1^{-1}\right)\mathrm{W}_1\,, \\
\bar{B}_{t,1} &= \sum_{i=1}^K m_i = \sum_{i=1}^K \hat{\Sigma}_{t,i}\hat{G}_{t,i}\hat{B}_{t,i} = \mathrm{W}_1^\top\Sigma_1^{-1}\sum_{i=1}^K\hat{\Sigma}_{t,i}\hat{G}_{t,i}\hat{B}_{t,i}\,.
\end{aligned}$$

This concludes the proof of the base case.

(II) Induction step. Let $\ell \in [L]/\{1\}$. Suppose that

$$\mathbb{P}\left(H_t \,|\, \psi_{*,\ell-1} = \psi_{\ell-1}\right) \propto \exp\left[-\frac{1}{2}\psi_{\ell-1}^\top\bar{G}_{t,\ell-1}\psi_{\ell-1} + \psi_{\ell-1}^\top\bar{B}_{t,\ell-1}\right]\,. \tag{27}$$

701 Then we want to show that

$$\mathbb{P}\left(H_t \,|\, \psi_{*,\ell} = \psi_\ell\right) \propto \exp\left[-\frac{1}{2}\psi_\ell^\top \bar{G}_{t,\ell}\psi_\ell + \psi_\ell^\top \bar{B}_{t,\ell}\right],$$

702 where

$$\begin{aligned}\bar{G}_{t,\ell} &= \mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell = \mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}(\Sigma_\ell^{-1} + \bar{G}_{t,\ell-1})^{-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\,,\\ \bar{B}_{t,\ell} &= \mathrm{W}_\ell^\top\Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\bar{B}_{t,\ell-1} = \mathrm{W}_\ell^\top\Sigma_\ell^{-1}(\Sigma_\ell^{-1} + \bar{G}_{t,\ell-1})^{-1}\bar{B}_{t,\ell-1}\,.\end{aligned}$$

703 To achieve this, we start by expressing $\mathbb{P}\left(H_t \,|\, \psi_{*,\ell} = \psi_\ell\right)$ in terms of $\mathbb{P}\left(H_t \,|\, \psi_{*,\ell-1} = \psi_{\ell-1}\right)$ as

$$\begin{aligned}&\mathbb{P}\left(H_t \,|\, \psi_{*,\ell} = \psi_\ell\right) = \int_{\psi_{\ell-1}} \mathbb{P}\left(H_t, \psi_{*,\ell-1} = \psi_{\ell-1} \,|\, \psi_{*,\ell} = \psi_\ell\right)\mathrm{d}\psi_{\ell-1}\,,\\ &= \int_{\psi_{\ell-1}} \mathbb{P}\left(H_t \,|\, \psi_{*,\ell-1} = \psi_{\ell-1}, \psi_{*,\ell} = \psi_\ell\right)\mathcal{N}(\psi_{\ell-1}; \mathrm{W}_\ell\psi_\ell, \Sigma_\ell)\,\mathrm{d}\psi_{\ell-1}\,,\\ &= \int_{\psi_{\ell-1}} \mathbb{P}\left(H_t \,|\, \psi_{*,\ell-1} = \psi_{\ell-1}\right)\mathcal{N}(\psi_{\ell-1}; \mathrm{W}_\ell\psi_\ell, \Sigma_\ell)\,\mathrm{d}\psi_{\ell-1}\,,\\ &\propto \int_{\psi_{\ell-1}} \exp\Big[-\frac{1}{2}\psi_{\ell-1}^\top\bar{G}_{t,\ell-1}\psi_{\ell-1} + \psi_{\ell-1}^\top\bar{B}_{t,\ell-1}\Big]\mathcal{N}(\psi_{\ell-1}; \mathrm{W}_\ell\psi_\ell, \Sigma_\ell)\,\mathrm{d}\psi_{\ell-1}\,,\\ &\propto \int_{\psi_{\ell-1}} \exp\Big[-\frac{1}{2}\psi_{\ell-1}^\top\bar{G}_{t,\ell-1}\psi_{\ell-1} + \psi_{\ell-1}^\top\bar{B}_{t,\ell-1}\\ &\qquad + (\psi_{\ell-1} - \mathrm{W}_\ell\psi_\ell)^\top\Lambda_\ell(\psi_{\ell-1} - \mathrm{W}_\ell\psi_\ell)\Big)\Big]\,\mathrm{d}\psi_{\ell-1}\,.\end{aligned}$$

704 Now let $S = \bar{G}_{t,\ell-1} + \Lambda_\ell$ and $V = \bar{B}_{t,\ell-1} + \Lambda_\ell\mathrm{W}_\ell\psi_\ell$. Then we have that,

$$\begin{aligned}&\mathbb{P}\left(H_t \,|\, \psi_{*,\ell} = \psi_\ell\right)\\ &\propto \int_{\psi_{\ell-1}} \exp\Big[-\frac{1}{2}\psi_{\ell-1}^\top\bar{G}_{t,\ell-1}\psi_{\ell-1} + \psi_{\ell-1}^\top\bar{B}_{t,\ell-1}\\ &\qquad + (\psi_{\ell-1} - \mathrm{W}_\ell\psi_\ell)^\top\Lambda_\ell(\psi_{\ell-1} - \mathrm{W}_\ell\psi_\ell)\Big)\Big]\,\mathrm{d}\psi_{\ell-1}\,,\\ &\propto \int_{\psi_{\ell-1}} \exp\Big[-\frac{1}{2}\Big(\psi_{\ell-1}^\top S\psi_{\ell-1} - 2\psi_{\ell-1}^\top\left(\bar{B}_{t,\ell-1} + \Lambda_\ell\mathrm{W}_\ell\psi_\ell\right) + \psi_\ell^\top\mathrm{W}_\ell^\top\Lambda_\ell\mathrm{W}_\ell\psi_\ell\Big)\Big]\,\mathrm{d}\psi_{\ell-1}\,,\\ &= \int_{\psi_{\ell-1}} \exp\Big[-\frac{1}{2}\Big(\psi_{\ell-1}^\top S(\psi_{\ell-1} - 2S^{-1}V) + \psi_\ell^\top\mathrm{W}_\ell^\top\Lambda_\ell\mathrm{W}_\ell\psi_\ell\Big)\Big]\,\mathrm{d}\psi_{\ell-1}\,,\\ &= \int_{\psi_{\ell-1}} \exp\Big[-\frac{1}{2}\Big((\psi_{\ell-1} - S^{-1}V)^\top S(\psi_{\ell-1} - S^{-1}V)\\ &\qquad + \psi_\ell^\top\mathrm{W}_\ell^\top\Lambda_\ell\mathrm{W}_\ell\psi_\ell - V^\top S^{-1}V\Big)\Big]\,\mathrm{d}\psi_{\ell-1}.\end{aligned}$$

705 In the second step, we omit constants in ψ_ℓ and $\psi_{\ell-1}$. Thus

$$\begin{aligned}&\mathbb{P}\left(H_t \,|\, \psi_{*,\ell} = \psi_\ell\right)\\ &\propto \int_{\psi_{\ell-1}}\exp\left[-\frac{1}{2}\left((\psi_{\ell-1} - S^{-1}V)^\top S(\psi_{\ell-1} - S^{-1}V) + \psi_\ell^\top\mathrm{W}_\ell^\top\Lambda_\ell\mathrm{W}_\ell\psi_\ell - V^\top S^{-1}V\right)\right]\mathrm{d}\psi_{\ell-1},\\ &\propto \exp\left[-\frac{1}{2}\left(\psi_\ell^\top\mathrm{W}_\ell^\top\Lambda_\ell\mathrm{W}_\ell\psi_\ell - V^\top S^{-1}V\right)\right].\end{aligned}$$

It follows that

$$
\begin{aligned}
&\mathbb{P}\left(H_t \,|\, \psi_{*,\ell} = \psi_\ell\right) \\
&\propto \exp\left[-\frac{1}{2}\left(\psi_\ell^\top \mathrm{W}_\ell^\top \Lambda_\ell \mathrm{W}_\ell \psi_\ell - V^\top S^{-1} V\right)\right], \\
&= \exp\left[-\frac{1}{2}\left(\psi_\ell^\top \mathrm{W}_\ell^\top \Lambda_\ell \mathrm{W}_\ell \psi_\ell - \left(\bar{B}_{t,\ell-1} + \Lambda_\ell \mathrm{W}_\ell \psi_\ell\right)^\top S^{-1}\left(\bar{B}_{t,\ell-1} + \Lambda_\ell \mathrm{W}_\ell \psi_\ell\right)\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\psi_\ell^\top \left(\mathrm{W}_\ell^\top \Lambda_\ell \mathrm{W}_\ell - \mathrm{W}_\ell^\top \Lambda_\ell S^{-1} \Lambda_\ell \mathrm{W}_\ell\right)\psi_\ell - 2\psi_\ell^\top \mathrm{W}_\ell^\top \Lambda_\ell S^{-1} \bar{B}_{t,\ell-1}\right)\right], \\
&= \exp\left[-\frac{1}{2}\psi_\ell^\top \bar{G}_{t,\ell}\psi_\ell + \psi_\ell^\top \bar{B}_{t,\ell}\right].
\end{aligned}
$$

In the last step, we omit constants in ψ_ℓ and we set

$$
\begin{aligned}
\bar{G}_{t,\ell} &= \mathrm{W}_\ell^\top\left(\Lambda_\ell - \Lambda_\ell S^{-1}\Lambda_\ell\right)\mathrm{W}_\ell = \mathrm{W}_\ell^\top\left(\Lambda_\ell - \Lambda_\ell(\Lambda_\ell + \bar{G}_{t,\ell-1})^{-1}\Sigma_\ell^{-1}\Lambda_\ell\right)\mathrm{W}_\ell\,, \\
\bar{B}_{t,\ell} &= \mathrm{W}_\ell^\top \Lambda_\ell S^{-1}\bar{B}_{t,\ell-1} = \mathrm{W}_\ell^\top \Lambda_\ell(\Lambda_\ell + \bar{G}_{t,\ell-1})^{-1}\bar{B}_{t,\ell-1}\,.
\end{aligned}
$$

This completes the proof. □

Similarly, this same proof applies when the reward distribution is not linear-Gaussian, with the approximation $\mathbb{P}\left(H_{t,i} \,|\, \theta_{*,i} = \theta\right) \approx \mathcal{N}\big(\theta; \hat{B}_{t,i}, \hat{G}_{t,i}^{-1}\big)$. Using this approximation in the derivations above leads to the same results.

C Posterior derivations for non-linear diffusion models

After deriving the posteriors for linear score functions f_ℓ, we now get back to the general case in (1), where the score functions are potentially non-linear. Approximation is needed since both the score functions and rewards can be non-linear. To avoid any computational challenges, we use a simple and intuitive approximation, where all posteriors $P_{t,i}$ and $Q_{t,\ell}$ are approximated by the Gaussian distributions in Appendix B.1, with few changes. First, the terms $\mathrm{W}_\ell\psi_\ell$ in (18) are replaced by $f_\ell(\psi_\ell)$. This accounts for the fact that the prior mean is now $f_\ell(\psi_\ell)$ rather than $\mathrm{W}_\ell\psi_\ell$, and this is the main difference between the linear diffusion model in (15) and the general, potentially non-linear, diffusion model in (1). Second, the matrix multiplications that involve the matrices W_ℓ in (20) and (21) are simply removed. Despite being simple, this approximation is efficient and avoids the computational burden of heavy approximate sampling algorithms required for each latent parameter. This is why deriving the exact posterior for linear score functions was key beyond enabling theoretical analyses. Moreover, this approximation retains some key attributes of exact posteriors. Specifically, in the absence of data, it recovers precisely the prior in (1), and as more data is accumulated, the influence of the prior diminishes.

D Regret proof and additional discussions

D.1 Sketch of the proof

We start with the following standard lemma upon which we build our analysis [Aouali et al., 2023b].

Lemma D.1. *Assume that* $\mathbb{P}\left(\theta_{*,i} = \theta \,|\, H_t\right) = \mathcal{N}(\theta; \check{\mu}_{t,i}, \check{\Sigma}_{t,i})$ *for any* $i \in [K]$*, then for any* $\delta \in (0, 1)$,

$$
\mathcal{BR}(n) \leq \sqrt{2n\log(1/\delta)}\sqrt{\mathbb{E}\left[\textstyle\sum_{t=1}^n \|X_t\|_{\check{\Sigma}_{t,A_t}}^2\right]} + cn\delta\,, \qquad \textit{where } c > 0 \textit{ is a constant}\,. \tag{28}
$$

Applying Lemma D.1 requires proving that the *marginal* action-posteriors $\mathbb{P}\left(\theta_{*,i} = \theta \,|\, H_t\right)$ in Eq. (3) are Gaussian and computing their covariances, while we only know the *conditional* action-posteriors $P_{t,i}$ and latent-posteriors $Q_{t,\ell}$. This is achieved by leveraging the preservation properties of the family of Gaussian distributions [Koller and Friedman, 2009] and the total covariance decomposition [Weiss, 2005] which leads to the next lemma.

Lemma D.2. *Let $t \in [n]$ and $i \in [K]$, then the marginal covariance matrix $\check{\Sigma}_{t,i}$ reads*

$$\check{\Sigma}_{t,i} = \hat{\Sigma}_{t,i} + \sum_{\ell \in [L]} \mathrm{P}_{i,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{i,\ell}^\top \,, \quad \text{where } \mathrm{P}_{i,\ell} = \hat{\Sigma}_{t,i} \Sigma_1^{-1} \mathrm{W}_1 \prod_{k=1}^{\ell-1} \bar{\Sigma}_{t,k} \Sigma_{k+1}^{-1} \mathrm{W}_{k+1} . \quad (29)$$

The marginal covariance matrix $\check{\Sigma}_{t,i}$ in Eq. (29) decomposes into $L + 1$ terms. The first term corresponds to the posterior uncertainty of $\theta_{*,i} \mid \psi_{*,1}$. The remaining L terms capture the posterior uncertainties of $\psi_{*,L}$ and $\psi_{*,\ell-1} \mid \psi_{*,\ell}$ for $\ell \in [L]/\{1\}$. These are then used to quantify the posterior information gain of latent parameters after one round as follows.

Lemma D.3 (Posterior information gain)**.** *Let $t \in [n]$ and $\ell \in [L]$, then*

$$\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} \succeq \sigma^{-2} \sigma_{\text{MAX}}^{-2\ell} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \,, \qquad \text{where } \sigma_{\text{MAX}}^2 = \max_{\ell \in [L+1]} 1 + \frac{\sigma_\ell^2}{\sigma^2} \,. \quad (30)$$

Finally, Lemma D.2 is used to decompose $\|X_t\|^2_{\check{\Sigma}_{t,A_t}}$ in Eq. (28) into $L + 1$ terms. Each term is bounded thanks to Lemma D.3. This results in the Bayes regret bound in Theorem 4.1.

D.2 Technical contributions

Our main technical contributions are the following.

Lemma D.2. In dTS, sampling is done hierarchically, meaning the marginal posterior distribution of $\theta_{*,i} | H_t$ is not explicitly defined. Instead, we use the conditional posterior distribution of $\theta_{*,i} | H_t, \psi_{*,1}$. The first contribution was deriving $\theta_{*,i} | H_t$ using the total covariance decomposition combined with an induction proof, as our posteriors in Section 3.1 were derived recursively. Unlike in Bayes regret analysis for standard Thompson sampling, where the posterior distribution of $\theta_{*,i} | H_t$ is predetermined due to the absence of latent parameters, our method necessitates this recursive total covariance decomposition, marking a first difference from the standard Bayesian proofs of Thompson sampling. Note that HierTS, which is developed for multi-task linear bandits, also employs total covariance decomposition, but it does so under the assumption of a single latent parameter; on which action parameters are centered. Our extension significantly differs as it is tailored for contextual bandits with multiple, successive levels of latent parameters, moving away from HierTS's assumption of a 1-level structure. Roughly speaking, HierTS when applied to contextual would consider a single-level hierarchy, where $\theta_{*,i} | \psi_{*,1} \sim \mathcal{N}(\psi_{*,1}, \Sigma_1)$ with $L = 1$. In contrast, our model proposes a multi-level hierarchy, where the first level is $\theta_{*,i} | \psi_{*,1} \sim \mathcal{N}(W_1 \psi_{*,1}, \Sigma_1)$. This also introduces a new aspect to our approach – the use of a linear function $W_1 \psi_{*,1}$, as opposed to HierTS's assumption where action parameters are centered directly on the latent parameter. Thus, while HierTS also uses the total covariance decomposition, our generalize it to multi-level hierarchies under L linear functions $W_\ell \psi_{*,\ell}$, instead of a single-level hierarchy under a single identity function $\psi_{*,1}$.

Lemma D.3. In Bayes regret proofs for standard Thompson sampling, we often quantify the posterior information gain. This is achieved by monitoring the increase in posterior precision for the action taken A_t in each round $t \in [n]$. However, in dTS, our analysis extends beyond this. We not only quantify the posterior information gain for the taken action but also for every latent parameter, since they are also learned. This lemma addresses this aspect. To elaborate, we use the recursive formulas in Section 3.1 that connect the posterior covariance of each latent parameter $\psi_{*,\ell}$ with the covariance of the posterior action parameters $\theta_{*,i}$. This allows us to propagate the information gain associated with the action taken in round A_t to all latent parameters $\psi_{*,\ell}$, for $\ell \in [L]$ by induction. This is a novel contribution, as it is not a feature of Bayes regret analyses in standard Thompson sampling.

Proposition 4.2. Building upon the insights of Theorem 4.1, we introduce the sparsity assumption **(A3)**. Under this assumption, we demonstrate that the Bayes regret outlined in Theorem 4.1 can be significantly refined. Specifically, the regret becomes contingent on dimensions $d_\ell \leq d$, as opposed to relying on the entire dimension d. This sparsity assumption is both a novel and a key technical contribution to our work. Its underlying principle is straightforward: the Bayes regret is influenced by the quantity of parameters that require learning. With the sparsity assumption, this number is reduced to less than d for each latent parameter. To substantiate this claim, we revisit the proof of Theorem 4.1 and modify a crucial equality. This adjustment results in a more precise representation by partitioning the covariance matrix of each latent parameter $\psi_{*,\ell}$ into blocks. These blocks comprise a $d_\ell \times d_\ell$ segment corresponding to the learnable d_ℓ parameters of $\psi_{*,\ell}$, and another block of size $(d - d_\ell) \times (d - d_\ell)$ that does not necessitate learning. This decomposition allows us to conclude that the final regret is solely dependent on d_ℓ, marking a significant refinement from the original theorem.

D.3 Proof of lemma D.2

In this proof, we heavily rely on the total covariance decomposition [Weiss, 2005]. Also, refer to [Hong et al., 2022b, Section 5.2] for a brief introduction to this decomposition. Now, from Eq. (17), we have that

$$\mathrm{cov}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] = \hat{\Sigma}_{t,i} = \left(\hat{G}_{t,i} + \Sigma_1^{-1}\right)^{-1}\,,$$
$$\mathbb{E}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] = \hat{\mu}_{t,i} = \hat{\Sigma}_{t,i}\left(\hat{G}_{t,i}\hat{B}_{t,i} + \Sigma_1^{-1}\mathrm{W}_1\psi_{*,1}\right)\,.$$

First, given H_t, $\mathrm{cov}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] = \left(\hat{G}_{t,i} + \Sigma_1^{-1}\right)^{-1}$ is constant. Thus

$$\mathbb{E}\left[\mathrm{cov}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] \,|\, H_t\right] = \mathrm{cov}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] = \left(\hat{G}_{t,i} + \Sigma_1^{-1}\right)^{-1} = \hat{\Sigma}_{t,i}\,.$$

In addition, given H_t, $\hat{\Sigma}_{t,i}$, $\hat{G}_{t,i}$ and $\hat{B}_{t,i}$ are constant. Thus

$$\begin{aligned}
\mathrm{cov}\left[\mathbb{E}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] \,|\, H_t\right] &= \mathrm{cov}\left[\hat{\Sigma}_{t,i}\left(\hat{G}_{t,i}\hat{B}_{t,i} + \Sigma_1^{-1}\mathrm{W}_1\psi_{*,1}\right) \,\Big|\, H_t\right]\,,\\
&= \mathrm{cov}\left[\hat{\Sigma}_{t,i}\Sigma_1^{-1}\mathrm{W}_1\psi_{*,1} \,\Big|\, H_t\right]\,,\\
&= \hat{\Sigma}_{t,i}\Sigma_1^{-1}\mathrm{W}_1\mathrm{cov}\left[\psi_{*,1} \,|\, H_t\right]\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,i}\,,\\
&= \hat{\Sigma}_{t,i}\Sigma_1^{-1}\mathrm{W}_1\bar{\bar{\Sigma}}_{t,1}\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,i}\,,
\end{aligned}$$

where $\bar{\bar{\Sigma}}_{t,1} = \mathrm{cov}\left[\psi_{*,1} \,|\, H_t\right]$ is the marginal posterior covariance of $\psi_{*,1}$. Finally, the total covariance decomposition [Weiss, 2005, Hong et al., 2022b] yields that

$$\begin{aligned}
\check{\Sigma}_{t,i} = \mathrm{cov}\left[\theta_{*,i} \,|\, H_t\right] &= \mathbb{E}\left[\mathrm{cov}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] \,|\, H_t\right] + \mathrm{cov}\left[\mathbb{E}\left[\theta_{*,i} \,|\, H_t, \psi_{*,1}\right] \,|\, H_t\right]\,,\\
&= \hat{\Sigma}_{t,i} + \hat{\Sigma}_{t,i}\Sigma_1^{-1}\mathrm{W}_1\bar{\bar{\Sigma}}_{t,1}\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,i}\,, \qquad (31)
\end{aligned}$$

However, $\bar{\bar{\Sigma}}_{t,1} = \mathrm{cov}\left[\psi_{*,1} \,|\, H_t\right]$ is different from $\bar{\Sigma}_{t,1} = \mathrm{cov}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right]$ that we already derived in Eq. (18). Thus we do not know the expression of $\bar{\bar{\Sigma}}_{t,1}$. But we can use the same total covariance decomposition trick to find it. Precisely, let $\bar{\bar{\Sigma}}_{t,\ell} = \mathrm{cov}\left[\psi_{*,\ell} \,|\, H_t\right]$ for any $\ell \in [L]$. Then we have that

$$\bar{\Sigma}_{t,1} = \mathrm{cov}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] = \left(\Sigma_2^{-1} + \bar{G}_{t,1}\right)^{-1}\,,$$
$$\bar{\mu}_{t,1} = \mathbb{E}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] = \bar{\Sigma}_{t,1}\left(\Sigma_2^{-1}\mathrm{W}_2\psi_{*,2} + \bar{B}_{t,1}\right)\,.$$

First, given H_t, $\mathrm{cov}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] = \left(\Sigma_2^{-1} + \bar{G}_{t,1}\right)^{-1}$ is constant. Thus

$$\mathbb{E}\left[\mathrm{cov}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] \,|\, H_t\right] = \mathrm{cov}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] = \bar{\Sigma}_{t,1}\,.$$

In addition, given H_t, $\bar{\Sigma}_{t,1}$, $\tilde{\Sigma}_{t,1}$ and $\bar{B}_{t,1}$ are constant. Thus

$$\begin{aligned}
\mathrm{cov}\left[\mathbb{E}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] \,|\, H_t\right] &= \mathrm{cov}\left[\bar{\Sigma}_{t,1}\left(\Sigma_2^{-1}\mathrm{W}_2\psi_{*,2} + \bar{B}_{t,1}\right) \,\Big|\, H_t\right]\,,\\
&= \mathrm{cov}\left[\bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\psi_{*,2} \,\big|\, H_t\right]\,,\\
&= \bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\mathrm{cov}\left[\psi_{*,2} \,|\, H_t\right]\mathrm{W}_2^\top\Sigma_2^{-1}\bar{\Sigma}_{t,1}\,,\\
&= \bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\bar{\bar{\Sigma}}_{t,2}\mathrm{W}_2^\top\Sigma_2^{-1}\bar{\Sigma}_{t,1}\,.
\end{aligned}$$

Finally, total covariance decomposition [Weiss, 2005, Hong et al., 2022b] leads to

$$\begin{aligned}
\bar{\bar{\Sigma}}_{t,1} = \mathrm{cov}\left[\psi_{*,1} \,|\, H_t\right] &= \mathbb{E}\left[\mathrm{cov}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] \,|\, H_t\right] + \mathrm{cov}\left[\mathbb{E}\left[\psi_{*,1} \,|\, H_t, \psi_{*,2}\right] \,|\, H_t\right]\,,\\
&= \bar{\Sigma}_{t,1} + \bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\bar{\bar{\Sigma}}_{t,2}\mathrm{W}_2^\top\Sigma_2^{-1}\bar{\Sigma}_{t,1}\,.
\end{aligned}$$

Now using the techniques, this can be generalized using the same technique as above to

$$\bar{\bar{\Sigma}}_{t,\ell} = \bar{\Sigma}_{t,\ell} + \bar{\Sigma}_{t,\ell}\Sigma_{\ell+1}^{-1}\mathrm{W}_{\ell+1}\bar{\bar{\Sigma}}_{t,\ell+1}\mathrm{W}_{\ell+1}^\top\Sigma_{\ell+1}^{-1}\bar{\Sigma}_{t,\ell}\,, \qquad \forall \ell \in [L-1]\,.$$

Then, by induction, we get that

$$\bar{\bar{\Sigma}}_{t,1} = \sum_{\ell \in [L]} \bar{\mathrm{P}}_\ell \bar{\Sigma}_{t,\ell} \bar{\mathrm{P}}_\ell^\top \,, \qquad\qquad \forall \ell \in [L-1] \,,$$

where we use that by definition $\bar{\bar{\Sigma}}_{t,L} = \mathrm{cov}\left[\psi_{*,L} \,|\, H_t\right] = \bar{\Sigma}_{t,L}$ and set $\bar{\mathrm{P}}_1 = I_d$ and $\bar{\mathrm{P}}_\ell = \prod_{k=1}^{\ell-1} \bar{\Sigma}_{t,k} \Sigma_{k+1}^{-1} \mathrm{W}_{k+1}$ for any $\ell \in [L]/\{1\}$. Plugging this in Eq. (31) leads to

$$\begin{aligned}
\check{\Sigma}_{t,i} &= \hat{\Sigma}_{t,i} + \sum_{\ell \in [L]} \hat{\Sigma}_{t,i} \Sigma_1^{-1} \mathrm{W}_1 \bar{\mathrm{P}}_\ell \bar{\Sigma}_{t,\ell} \bar{\mathrm{P}}_\ell^\top \mathrm{W}_1^\top \Sigma_1^{-1} \hat{\Sigma}_{t,i} \,, \\
&= \hat{\Sigma}_{t,i} + \sum_{\ell \in [L]} \hat{\Sigma}_{t,i} \Sigma_1^{-1} \mathrm{W}_1 \bar{\mathrm{P}}_\ell \bar{\Sigma}_{t,\ell} (\hat{\Sigma}_{t,i} \Sigma_1^{-1} \mathrm{W}_1)^\top \,, \\
&= \hat{\Sigma}_{t,i} + \sum_{\ell \in [L]} \mathrm{P}_{i,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{i,\ell}^\top \,,
\end{aligned}$$

where $\mathrm{P}_{i,\ell} = \hat{\Sigma}_{t,i} \Sigma_1^{-1} \mathrm{W}_1 \bar{\mathrm{P}}_\ell = \hat{\Sigma}_{t,i} \Sigma_1^{-1} \mathrm{W}_1 \prod_{k=1}^{\ell-1} \bar{\Sigma}_{t,k} \Sigma_{k+1}^{-1} \mathrm{W}_{k+1}$.

D.4 Proof of lemma D.3

We prove this result by induction. We start with the base case when $\ell = 1$.

(I) Base case. Let $u = \sigma^{-1} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t$ From the expression of $\bar{\Sigma}_{t,1}$ in Eq. (18), we have that

$$\begin{aligned}
\bar{\Sigma}_{t+1,1}^{-1} - \bar{\Sigma}_{t,1}^{-1} &= \mathrm{W}_1^\top \left(\Sigma_1^{-1} - \Sigma_1^{-1} (\hat{\Sigma}_{t,A_t}^{-1} + \sigma^{-2} X_t X_t^\top)^{-1} \Sigma_1^{-1} - (\Sigma_1^{-1} - \Sigma_1^{-1} \hat{\Sigma}_{t,A_t} \Sigma_1^{-1}) \right) \mathrm{W}_1 \,, \\
&= \mathrm{W}_1^\top \left(\Sigma_1^{-1} (\hat{\Sigma}_{t,A_t} - (\hat{\Sigma}_{t,A_t}^{-1} + \sigma^{-2} X_t X_t^\top)^{-1}) \Sigma_1^{-1} \right) \mathrm{W}_1 \,, \\
&= \mathrm{W}_1^\top \left(\Sigma_1^{-1} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} (I_d - (I_d + \sigma^{-2} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \hat{\Sigma}_{t,A_t}^{\frac{1}{2}})^{-1}) \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} \Sigma_1^{-1} \right) \mathrm{W}_1 \,, \\
&= \mathrm{W}_1^\top \left(\Sigma_1^{-1} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} (I_d - (I_d + u u^\top)^{-1}) \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} \Sigma_1^{-1} \right) \mathrm{W}_1 \,, \\
&\overset{(i)}{=} \mathrm{W}_1^\top \left(\Sigma_1^{-1} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} \frac{u u^\top}{1 + u^\top u} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} \Sigma_1^{-1} \right) \mathrm{W}_1 \,, \\
&\overset{(ii)}{=} \sigma^{-2} \mathrm{W}_1^\top \Sigma_1^{-1} \hat{\Sigma}_{t,A_t} \frac{X_t X_t^\top}{1 + u^\top u} \hat{\Sigma}_{t,A_t} \Sigma_1^{-1} \mathrm{W}_1 \,.
\end{aligned} \tag{32}$$

In (i) we use the Sherman-Morrison formula. Note that (ii) says that $\bar{\Sigma}_{t+1,1}^{-1} - \bar{\Sigma}_{t,1}^{-1}$ is one-rank which we will also need in induction step. Now, we have that $\|X_t\|^2 = 1$. Therefore,

$$1 + u^\top u = 1 + \sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t \leq 1 + \sigma^{-2} \lambda_1(\Sigma_1) \|X_t\|^2 = 1 + \sigma^{-2} \sigma_1^2 \leq \sigma_{\text{MAX}}^2 \,,$$

where we use that by definition of σ_{MAX}^2 in Lemma D.3, we have that $\sigma_{\text{MAX}}^2 \geq 1 + \sigma^{-2} \sigma_1^2$. Therefore, by taking the inverse, we get that $\frac{1}{1 + u^\top u} \geq \sigma_{\text{MAX}}^{-2}$. Combining this with Eq. (32) leads to

$$\bar{\Sigma}_{t+1,1}^{-1} - \bar{\Sigma}_{t,1}^{-1} \succeq \sigma^{-2} \sigma_{\text{MAX}}^{-2} \mathrm{W}_1^\top \Sigma_1^{-1} \hat{\Sigma}_{t,A_t} X_t X_t^\top \hat{\Sigma}_{t,A_t} \Sigma_1^{-1} \mathrm{W}_1$$

Noticing that $\mathrm{P}_{A_t,1} = \hat{\Sigma}_{t,A_t} \Sigma_1^{-1} \mathrm{W}_1$ concludes the proof of the base case when $\ell = 1$.

(II) Induction step. Let $\ell \in [L]/\{1\}$ and suppose that $\bar{\Sigma}_{t+1,\ell-1}^{-1} - \bar{\Sigma}_{t,\ell-1}^{-1}$ is one-rank and that it holds for $\ell - 1$ that

$$\bar{\Sigma}_{t+1,\ell-1}^{-1} - \bar{\Sigma}_{t,\ell-1}^{-1} \succeq \sigma^{-2} \sigma_{\text{MAX}}^{-2(\ell-1)} \mathrm{P}_{A_t,\ell-1}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell-1} \,, \quad \text{where } \sigma_{\text{MAX}}^{-2} = \max_{\ell \in [L]} 1 + \sigma^{-2} \sigma_\ell^2 .$$

Then, we want to show that $\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1}$ is also one-rank and that it holds that

$$\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} \succeq \sigma^{-2} \sigma_{\text{MAX}}^{-2\ell} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \,, \qquad \text{where } \sigma_{\text{MAX}}^{-2} = \max_{\ell \in [L]} 1 + \sigma^{-2} \sigma_\ell^2 .$$

This is achieved as follows. First, we notice that by the induction hypothesis, we have that $\tilde{\Sigma}_{t+1,\ell-1}^{-1} - \bar{G}_{t,\ell-1} = \bar{\Sigma}_{t+1,\ell-1}^{-1} - \bar{\Sigma}_{t,\ell-1}^{-1}$ is one-rank. In addition, the matrix is positive semi-definite. Thus we can write it as $\tilde{\Sigma}_{t+1,\ell-1}^{-1} - \bar{G}_{t,\ell-1} = uu^\top$ where $u \in \mathbb{R}^d$. Then, similarly to the base case, we have

$$
\begin{aligned}
\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} &= \tilde{\Sigma}_{t+1,\ell}^{-1} - \tilde{\Sigma}_{t,\ell}^{-1}\,,\\
&= \mathrm{W}_\ell^\top \left(\Sigma_\ell + \tilde{\Sigma}_{t+1,\ell-1}\right)^{-1}\mathrm{W}_\ell - \mathrm{W}_\ell^\top \left(\Sigma_\ell + \tilde{\Sigma}_{t,\ell-1}\right)^{-1}\mathrm{W}_\ell\,,\\
&= \mathrm{W}_\ell^\top \left[\left(\Sigma_\ell + \tilde{\Sigma}_{t+1,\ell-1}\right)^{-1} - \left(\Sigma_\ell + \tilde{\Sigma}_{t,\ell-1}\right)^{-1}\right]\mathrm{W}_\ell\,,\\
&= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\left[\left(\Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}\right)^{-1} - \left(\Sigma_\ell^{-1} + \tilde{\Sigma}_{t+1,\ell-1}^{-1}\right)^{-1}\right]\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\left[\left(\Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}\right)^{-1} - \left(\Sigma_\ell^{-1} + \bar{G}_{t,\ell-1} + \tilde{\Sigma}_{t+1,\ell-1}^{-1} - \bar{G}_{t,\ell-1}\right)^{-1}\right]\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\left[\left(\Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}\right)^{-1} - \left(\Sigma_\ell^{-1} + \bar{G}_{t,\ell-1} + uu^\top\right)^{-1}\right]\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\left[\bar{\Sigma}_{t,\ell-1} - \left(\bar{\Sigma}_{t,\ell-1}^{-1} + uu^\top\right)^{-1}\right]\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\left[\bar{\Sigma}_{t,\ell-1}\frac{uu^\top}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\bar{\Sigma}_{t,\ell-1}\right]\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\frac{uu^\top}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\mathrm{W}_\ell
\end{aligned}
$$

However, we it follows from the induction hypothesis that $uu^\top = \tilde{\Sigma}_{t+1,\ell-1}^{-1} - \bar{G}_{t,\ell-1} = \bar{\Sigma}_{t+1,\ell-1}^{-1} - \bar{\Sigma}_{t,\ell-1}^{-1} \succeq \sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}\mathrm{P}_{A_t,\ell-1}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell-1}$. Therefore,

$$
\begin{aligned}
\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} &= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\frac{uu^\top}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&\succeq \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}\mathrm{P}_{A_t,\ell-1}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell-1}}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&= \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\mathrm{W}_\ell^\top \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\mathrm{P}_{A_t,\ell-1}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\mathrm{W}_\ell\,,\\
&= \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}\,.
\end{aligned}
$$

Finally, we use that $1 + u^\top\bar{\Sigma}_{t,\ell-1}u \leq 1 + \|u\|_2\lambda_1(\bar{\Sigma}_{t,\ell-1}) \leq 1 + \sigma^{-2}\sigma_\ell^2$. Here we use that $\|u\|_2 \leq \sigma^{-2}$, which can also be proven by induction, and that $\lambda_1(\bar{\Sigma}_{t,\ell-1}) \leq \sigma_\ell^2$, which follows from the expression of $\bar{\Sigma}_{t,\ell-1}$ in Section 3.1. Therefore, we have that

$$
\begin{aligned}
\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} &\succeq \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}\,,\\
&\succeq \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + \sigma^{-2}\sigma_\ell^2}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}\,,\\
&\succeq \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}\,,
\end{aligned}
$$

where the last inequality follows from the definition of $\sigma_{\text{MAX}}^2 = \max_{\ell\in[L]} 1 + \sigma^{-2}\sigma_\ell^2$. This concludes the proof.

D.5 Proof of theorem 4.1

We start with the following standard result which we borrow from [Hong et al., 2022a, Aouali et al., 2023b],

$$
\mathcal{BR}(n) \leq \sqrt{2n\log(1/\delta)}\sqrt{\mathbb{E}\left[\sum_{t=1}^{n}\|X_t\|_{\hat{\Sigma}_{t,A_t}}^2\right]} + cn\delta\,, \qquad \text{where } c > 0 \text{ is a constant}\,. \tag{33}
$$

Then we use Lemma D.2 and express the marginal covariance $\check{\Sigma}_{t,A_t}$ as

$$\check{\Sigma}_{t,i} = \hat{\Sigma}_{t,i} + \sum_{\ell \in [L]} \mathrm{P}_{i,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{i,\ell}^\top , \qquad \text{where } \mathrm{P}_{i,\ell} = \hat{\Sigma}_{t,i} \Sigma_1^{-1} \mathrm{W}_1 \prod_{k=1}^{\ell-1} \bar{\Sigma}_{t,k} \Sigma_{k+1}^{-1} \mathrm{W}_{k+1} . \tag{34}$$

Therefore, we can decompose $\|X_t\|^2_{\check{\Sigma}_{t,A_t}}$ as

$$\begin{aligned} \|X_t\|^2_{\check{\Sigma}_{t,A_t}} &= \sigma^2 \frac{X_t^\top \check{\Sigma}_{t,A_t} X_t}{\sigma^2} \overset{(i)}{=} \sigma^2 \big(\sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t + \sigma^{-2} \sum_{\ell \in [L]} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t \big) , \\ &\overset{(ii)}{\leq} c_0 \log(1 + \sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t) + \sum_{\ell \in [L]} c_\ell \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) , \end{aligned} \tag{35}$$

where (i) follows from Eq. (34), and we use the following inequality in (ii)

$$x = \frac{x}{\log(1+x)} \log(1+x) \leq \left(\max_{x \in [0,u]} \frac{x}{\log(1+x)} \right) \log(1+x) = \frac{u}{\log(1+u)} \log(1+x) ,$$

which holds for any $x \in [0, u]$, where constants c_0 and c_ℓ are derived as

$$c_0 = \frac{\sigma_1^2}{\log(1 + \frac{\sigma_1^2}{\sigma^2})} , \quad c_\ell = \frac{\sigma_{\ell+1}^2}{\log(1 + \frac{\sigma_{\ell+1}^2}{\sigma^2})} , \text{with the convention that } \sigma_{L+1} = 1 .$$

The derivation of c_0 uses that

$$X_t^\top \hat{\Sigma}_{t,A_t} X_t \leq \lambda_1(\hat{\Sigma}_{t,A_t}) \|X_t\|^2 \leq \lambda_d^{-1}(\Sigma_1^{-1} + G_{t,A_t}) \leq \lambda_d^{-1}(\Sigma_1^{-1}) = \lambda_1(\Sigma_1) = \sigma_1^2 .$$

The derivation of c_ℓ follows from

$$X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t \leq \lambda_1(\mathrm{P}_{A_t,\ell} \mathrm{P}_{A_t,\ell}^\top) \lambda_1(\bar{\Sigma}_{t,\ell}) \|X_t\|^2 \leq \sigma_{\ell+1}^2 .$$

Therefore, from Eq. (35) and Eq. (33), we get that

$$\begin{aligned} \mathcal{BR}(n) \leq \sqrt{2n \log(1/\delta)} \Big(\mathbb{E}\Big[c_0 \sum_{t=1}^n \log(1 + \sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t) \\ + \sum_{\ell \in [L]} c_\ell \sum_{t=1}^n \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) \Big] \Big)^{\frac{1}{2}} + cn\delta \end{aligned} \tag{36}$$

Now we focus on bounding the logarithmic terms in Eq. (36).

(I) First term in Eq. (36) We first rewrite this term as

$$\begin{aligned} \log(1 + \sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t) &\overset{(i)}{=} \log\det(I_d + \sigma^{-2} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \hat{\Sigma}_{t,A_t}^{\frac{1}{2}}) , \\ &= \log\det(\hat{\Sigma}_{t,A_t}^{-1} + \sigma^{-2} X_t X_t^\top) - \log\det(\hat{\Sigma}_{t,A_t}^{-1}) = \log\det(\hat{\Sigma}_{t+1,A_t}^{-1}) - \log\det(\hat{\Sigma}_{t,A_t}^{-1}) , \end{aligned}$$

where (i) follows from the Weinstein–Aronszajn identity. Then we sum over all rounds $t \in [n]$, and get a telescoping

$$\begin{aligned} \sum_{t=1}^n \log\det(I_d + \sigma^{-2} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \hat{\Sigma}_{t,A_t}^{\frac{1}{2}}) &= \sum_{t=1}^n \log\det(\hat{\Sigma}_{t+1,A_t}^{-1}) - \log\det(\hat{\Sigma}_{t,A_t}^{-1}) , \\ &= \sum_{t=1}^n \sum_{i=1}^K \log\det(\hat{\Sigma}_{t+1,i}^{-1}) - \log\det(\hat{\Sigma}_{t,i}^{-1}) = \sum_{i=1}^K \sum_{t=1}^n \log\det(\hat{\Sigma}_{t+1,i}^{-1}) - \log\det(\hat{\Sigma}_{t,i}^{-1}) , \\ &= \sum_{i=1}^K \log\det(\hat{\Sigma}_{n+1,i}^{-1}) - \log\det(\hat{\Sigma}_{1,i}^{-1}) \overset{(i)}{=} \sum_{i=1}^K \log\det(\Sigma_1^{\frac{1}{2}} \hat{\Sigma}_{n+1,i}^{-1} \Sigma_1^{\frac{1}{2}}) , \end{aligned}$$

where (i) follows from the fact that $\hat{\Sigma}_{1,i} = \Sigma_1$. Now we use the inequality of arithmetic and geometric means and get

$$
\begin{aligned}
\sum_{t=1}^{n} \log\det(I_d + \sigma^{-2}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \hat{\Sigma}_{t,A_t}^{\frac{1}{2}}) &= \sum_{i=1}^{K} \log\det(\Sigma_1^{\frac{1}{2}}\hat{\Sigma}_{n+1,i}^{-1}\Sigma_1^{\frac{1}{2}})\,, \\
&\leq \sum_{i=1}^{K} d\log\left(\frac{1}{d}\operatorname{Tr}(\Sigma_1^{\frac{1}{2}}\hat{\Sigma}_{n+1,i}^{-1}\Sigma_1^{\frac{1}{2}})\right)\,, \\
&\leq \sum_{i=1}^{K} d\log\left(1 + \frac{n}{d}\frac{\sigma_1^2}{\sigma^2}\right) = Kd\log\left(1 + \frac{n}{d}\frac{\sigma_1^2}{\sigma^2}\right)\,.
\end{aligned}
\tag{37}
$$

(II) Remaining terms in Eq. (36) Let $\ell \in [L]$. Then we have that

$$
\begin{aligned}
\log(1 + \sigma^{-2}X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t) &= \sigma_{\text{MAX}}^{2\ell}\sigma_{\text{MAX}}^{-2\ell}\log(1 + \sigma^{-2}X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t)\,, \\
&\leq \sigma_{\text{MAX}}^{2\ell}\log(1 + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell}X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t)\,, \\
&\overset{(i)}{=} \sigma_{\text{MAX}}^{2\ell}\log\det(I_d + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell}\bar{\Sigma}_{t,\ell}^{\frac{1}{2}}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}^{\frac{1}{2}})\,, \\
&= \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{t,\ell}^{-1} + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}) - \log\det(\bar{\Sigma}_{t,\ell}^{-1})\Big)\,,
\end{aligned}
$$

where we use the Weinstein–Aronszajn identity in (i). Now we know from Lemma D.3 that the following inequality holds $\sigma^{-2}\sigma_{\text{MAX}}^{-2\ell}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \preceq \bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1}$. As a result, we get that $\bar{\Sigma}_{t,\ell}^{-1} + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell}\mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \preceq \bar{\Sigma}_{t+1,\ell}^{-1}$. Thus,

$$
\log(1 + \sigma^{-2}X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t) \leq \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{t+1,\ell}^{-1}) - \log\det(\bar{\Sigma}_{t,\ell}^{-1})\Big)\,,
$$

Then we sum over all rounds $t \in [n]$, and get a telescoping

$$
\begin{aligned}
\sum_{t=1}^{n}\log(1 + \sigma^{-2}X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t) &\leq \sigma_{\text{MAX}}^{2\ell}\sum_{t=1}^{n}\log\det(\bar{\Sigma}_{t+1,\ell}^{-1}) - \log\det(\bar{\Sigma}_{t,\ell}^{-1})\,, \\
&= \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{n+1,\ell}^{-1}) - \log\det(\bar{\Sigma}_{1,\ell}^{-1})\Big)\,, \\
&\overset{(i)}{=} \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{n+1,\ell}^{-1}) - \log\det(\Sigma_{\ell+1}^{-1})\Big)\,, \\
&= \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})\Big)\,,
\end{aligned}
$$

where we use that $\bar{\Sigma}_{1,\ell} = \Sigma_{\ell+1}$ in (i). Finally, we use the inequality of arithmetic and geometric means and get that

$$
\begin{aligned}
\sum_{t=1}^{n}\log(1 + \sigma^{-2}X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t) &\leq \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})\Big)\,, \\
&\leq d\sigma_{\text{MAX}}^{2\ell}\log\left(\frac{1}{d}\operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})\right)\,, \\
&\leq d\sigma_{\text{MAX}}^{2\ell}\log\left(1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\right)\,,
\end{aligned}
\tag{38}
$$

The last inequality follows from the expression of $\bar{\Sigma}_{n+1,\ell}^{-1}$ in Eq. (18) that leads to

$$
\begin{aligned}
\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}} &= I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\bar{G}_{t,\ell}\Sigma_{\ell+1}^{\frac{1}{2}}\,, \\
&= I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}}\,,
\end{aligned}
\tag{39}
$$

since $\bar{G}_{t,\ell} = \mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell$. This allows us to bound $\frac{1}{d}\operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})$ as

$$\begin{aligned}
\frac{1}{d}\operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}}) &= \frac{1}{d}\operatorname{Tr}(I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}})\,,\\
&= \frac{1}{d}(d + \operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}})\,,\\
&\leq 1 + \frac{1}{d}\sum_{k=1}^{d}\lambda_1(\Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}}\,,\\
&\leq 1 + \frac{1}{d}\sum_{k=1}^{d}\lambda_1(\Sigma_{\ell+1})\lambda_1(\mathrm{W}_\ell^\top\mathrm{W}_\ell)\lambda_1\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\,,\\
&\leq 1 + \frac{1}{d}\sum_{k=1}^{d}\lambda_1(\Sigma_{\ell+1})\lambda_1(\mathrm{W}_\ell^\top\mathrm{W}_\ell)\lambda_1\big(\Sigma_\ell^{-1}\big)\,,\\
&\leq 1 + \frac{1}{d}\sum_{k=1}^{d}\frac{\sigma_{\ell+1}^2}{\sigma_\ell^2} = 1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\,,
\end{aligned} \tag{40}$$

where we use the assumption that $\lambda_1(\mathrm{W}_\ell^\top\mathrm{W}_\ell) = 1$ **(A2)** and that $\lambda_1(\Sigma_{\ell+1}) = \sigma_{\ell+1}^2$ and $\lambda_1(\Sigma_\ell^{-1}) = 1/\sigma_\ell^2$. This is because $\Sigma_\ell = \sigma_\ell^2 I_d$ for any $\ell \in [L+1]$. Finally, plugging Eqs. (37) and (38) in Eq. (36) concludes the proof.

D.6 Proof of proposition 4.2

We use exactly the same proof in Appendix D.5, with one change to account for the sparsity assumption **(A3)**. The change corresponds to Eq. (38). First, recall that Eq. (38) writes

$$\sum_{t=1}^{n}\log(1 + \sigma^{-2}X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t) \leq \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})\Big)\,,$$

where

$$\begin{aligned}
\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}} &= I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}}\,,\\
&= I_d + \sigma_{\ell+1}^2\mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\,,
\end{aligned} \tag{41}$$

where the second equality follows from the assumption that $\Sigma_{\ell+1} = \sigma_{\ell+1}^2 I_d$. But notice that in our assumption, **(A3)**, we assume that $\mathrm{W}_\ell = (\bar{\mathrm{W}}_\ell, 0_{d,d-d_\ell})$, where $\bar{\mathrm{W}}_\ell \in \mathbb{R}^{d\times d_\ell}$ for any $\ell \in [L]$. Therefore, we have that for any $d\times d$ matrix $\mathrm{B} \in \mathbb{R}^{dd\times d}$, the following holds, $\mathrm{W}_\ell^\top \mathrm{B}\mathrm{W}_\ell = \begin{pmatrix}\bar{\mathrm{W}}_\ell^\top\mathrm{B}\bar{\mathrm{W}}_\ell & 0_{d_\ell,d-d_\ell}\\ 0_{d-d_\ell,d_\ell} & 0_{d-d_\ell,d-d_\ell}\end{pmatrix}$. In particular, we have that

$$\mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell = \begin{pmatrix}\bar{\mathrm{W}}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\bar{\mathrm{W}}_\ell & 0_{d_\ell,d-d_\ell}\\ 0_{d-d_\ell,d_\ell} & 0_{d-d_\ell,d-d_\ell}\end{pmatrix}. \tag{42}$$

Therefore, plugging this in Eq. (41) yields that

$$\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}} = \begin{pmatrix}I_{d_\ell} + \sigma_{\ell+1}^2\bar{\mathrm{W}}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\bar{\mathrm{W}}_\ell & 0_{d_\ell,d-d_\ell}\\ 0_{d-d_\ell,d_\ell} & I_{d-d_\ell}\end{pmatrix}. \tag{43}$$

As a result, $\det(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{n+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}}) = \det(I_{d_\ell} + \sigma_{\ell+1}^2\bar{\mathrm{W}}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\bar{\mathrm{W}}_\ell)$. This allows us to move the problem from a d-dimensional one to a d_ℓ-dimensional one. Then we use the inequality of arithmetic and geometric means and get that

$$\sum_{t=1}^{n} \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) \leq \sigma_{\text{MAX}}^{2\ell} \Big(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}} \bar{\Sigma}_{n+1,\ell}^{-1} \Sigma_{\ell+1}^{\frac{1}{2}}) \Big)\,,$$
$$= \sigma_{\text{MAX}}^{2\ell} \log\det(I_{d_\ell} + \sigma_{\ell+1}^2 \bar{\mathrm{W}}_\ell^\top (\Sigma_\ell^{-1} - \Sigma_\ell^{-1} \bar{\Sigma}_{t,\ell-1} \Sigma_\ell^{-1}) \bar{\mathrm{W}}_\ell)\,,$$
$$\leq d_\ell \sigma_{\text{MAX}}^{2\ell} \log\left(\frac{1}{d_\ell} \mathrm{Tr}(I_{d_\ell} + \sigma_{\ell+1}^2 \bar{\mathrm{W}}_\ell^\top (\Sigma_\ell^{-1} - \Sigma_\ell^{-1} \bar{\Sigma}_{t,\ell-1} \Sigma_\ell^{-1}) \bar{\mathrm{W}}_\ell) \right)\,,$$
$$\leq d_\ell \sigma_{\text{MAX}}^{2\ell} \log\left(1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2} \right)\,. \tag{44}$$

To get the last inequality, we use derivations similar to the ones we used in Eq. (40). Finally, the desired result in obtained by replacing Eq. (38) by Eq. (44) in the previous proof in Appendix D.5.

D.7 Additional discussion: link to two-level hierarchies

The linear diffusion (15) can be marginalized into a 2-level hierarchy using two different strategies. The first one yields,

$$\begin{aligned} \psi_{*,L} &\sim \mathcal{N}(0, \sigma_{L+1}^2 \mathrm{B}_L \mathrm{B}_L^\top)\,, \\ \theta_{*,i} \mid \psi_{*,L} &\sim \mathcal{N}(\psi_{*,L}, \Omega_1)\,, \qquad \forall i \in [K]\,, \end{aligned} \tag{45}$$

with $\Omega_1 = \sigma_1^2 I_d + \sum_{\ell=1}^{L-1} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top$ and $\mathrm{B}_\ell = \prod_{k=1}^{\ell} \mathrm{W}_k$. The second strategy yields,

$$\begin{aligned} \psi_{*,1} &\sim \mathcal{N}(0, \Omega_2)\,, \\ \theta_{*,i} \mid \psi_{*,1} &\sim \mathcal{N}(\psi_{*,1}, \sigma_1^2 I_d)\,, \qquad \forall i \in [K]\,, \end{aligned} \tag{46}$$

where $\Omega_2 = \sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top$. Recently, HierTS [Hong et al., 2022b] was developed for such two-level graphical models, and we call HierTS under (45) by HierTS-1 and HierTS under (46) by HierTS-2. Then, we start by highlighting the differences between these two variants of HierTS. First, their regret bounds scale as

$$\texttt{HierTS-1} : \tilde{\mathcal{O}}\big(\sqrt{nd(K \textstyle\sum_{\ell=1}^{L} \sigma_\ell^2 + L\sigma_{L+1}^2})\big)\,, \quad \texttt{HierTS-2} : \tilde{\mathcal{O}}\big(\sqrt{nd(K\sigma_1^2 + \textstyle\sum_{\ell=1}^{L} \sigma_{\ell+1}^2)}\big)\,.$$

When $K \approx L$, the regret bounds of HierTS-1 and HierTS-2 are similar. However, when $K > L$, HierTS-2 outperforms HierTS-1. This is because HierTS-2 puts more uncertainty on a single d-dimensional latent parameter $\psi_{*,1}$, rather than K individual d-dimensional action parameters $\theta_{*,i}$. More importantly, HierTS-1 implicitly assumes that action parameters $\theta_{*,i}$ are conditionally independent given $\psi_{*,L}$, which is not true. Consequently, HierTS-2 outperforms HierTS-1. Note that, under the linear diffusion model (15), dTS and HierTS-2 have roughly similar regret bounds. Specifically, their regret bounds dependency on K is identical, where both methods involve multiplying K by σ_1^2, and both enjoy improved performance compared to HierTS-1. That said, note that Theorem 4.1 and Proposition 4.2 provide an understanding of how dTS's regret scales under linear score functions f_ℓ, and do not say that using dTS is better than using HierTS when the score functions f_ℓ are linear since the latter can be obtained by a proper marginalization of latent parameters (i.e., HierTS-2 instead of HierTS-1). While such a comparison is not the goal of this work, we still provide it for completeness next.

When the mixing matrices W_ℓ are dense (i.e., assumption **(A3)** is not applicable), dTS and HierTS-2 have comparable regret bounds and computational efficiency. However, under the sparsity assumption **(A3)** and with mixing matrices that allow for conditional independence of $\psi_{*,1}$ coordinates given $\psi_{*,2}$, dTS enjoys a computational advantage over HierTS-2. This advantage explains why works focusing on multi-level hierarchies typically benchmark their algorithms against two-level structures akin to HierTS-1, rather than the more competitive HierTS-2. This is also consistent with prior works in Bayesian bandits using multi-level hierarchies, such as Tree-based priors [Hong et al., 2022a], which compared their method to HierTS-1. In line with this, we also compared dTS with HierTS-1 in our experiments. But this is only given for completeness as this is not the aim of Theorem 4.1 and Proposition 4.2. More importantly, HierTS is inapplicable in the general case in (1) with non-linear score functions since the latent parameters cannot be analytically marginalized.

E Broader impact

This work contributes to the development and analysis of practical algorithms for online learning to act under uncertainty. While our generic setting and algorithms have broad potential applications, the specific downstream social impacts are inherently dependent on the chosen application domain. Nevertheless, we acknowledge the crucial need to consider potential biases that may be present in pre-trained diffusion models, given that our method relies on them.

F Limitations

Our work investigated contextual bandits, laying the groundwork for future exploration into reinforcement learning. This exploration can be done from both practical (empirical) and theoretical angles. While our method, which approximates rewards using a Gaussian distribution, worked well for linear rewards and those following a generalized linear model, its effectiveness in real-world, complex scenarios needs further testing. Another interesting direction for future research is pre-training the diffusion model prior. Hsieh et al. [2023] proposed a method for this in multi-armed bandits, but its application to contextual bandits remains unexplored.

G Amount of computation required

Our experiments were conducted on internal machines with 30 CPUs and thus they required a moderate amount of computation. These experiments are also reproducible with minimal computational resources.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are supported by the theory in Section 4 (with proofs provided in the appendix) and experiments in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations were discussed in Section 6 and Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Assumptions are mentioned in the main text. Complete proofs are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Information needed to reproduce the main experimental results of the paper is described in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code for the main experiments is shared in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental details are described in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Standard error bars are included in the figures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As mentioned in Appendix G, our experiments were conducted on internal machines with 30 CPUs and thus they required a moderate amount of computation. These experiments are also reproducible with minimal computational resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work contributes to the development and theoretical analysis of online learning to act under uncertainty and it adheres to the Neurips Code Of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader Impacts are discussed in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper is mainly theoretical and the used data is simulated. Thus, we believe that our work poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: To the best of our knowledge, all relevant and used papers were cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include our code as supplementary material, with all details needed for reproducibility given in Section 5.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.