Deep Hierarchy in Bandits

Joey Hong 1 Branislav Kveton 2 Sumeet Katariya 2 Manzil Zaheer 3 Mohammad Ghavamzadeh 4

Abstract
Mean rewards of actions are often correlated. The form of these correlations may be complex and unknown a priori, such as the preferences of users for recommended products and their categories. To maximize statistical efficiency, it is important to leverage these correlations when learning. We formulate a bandit variant of this problem where the correlations of mean action rewards are represented by a hierarchical Bayesian model with latent variables. Since the hierarchy can have multiple layers, we call it deep. We propose a hierarchical Thompson sampling algorithm (HierTS) for this problem and show how to implement it efficiently for Gaussian hierarchies. The efficient implementation is possible due to a novel exact hierarchical representation of the posterior, which itself is of independent interest. We use this exact posterior to analyze the Bayes regret of HierTS. Our regret bounds reflect the structure of the problem, that the regret decreases with more informative priors, and can be recast to highlight reduced dependence on the number of actions. We confirm these theoretical findings empirically, in both synthetic and real-world experiments.

1. Introduction
A contextual bandit (Li et al., 2010; Chu et al., 2011) is a sequential decision-making problem where a learning agent sequentially interacts with an environment over $n$ rounds. In each round, the agent observes a context, chooses one of $K$ possible actions, and then receives a reward for the taken action. The agent aims to maximize its expected cumulative reward over $n$ rounds. It does not know the mean rewards of the actions $a$ priori and learns them by taking the actions. This forces the agent to choose between exploring actions to learn about them and exploiting the action with the highest estimated reward. As an example, in online shopping, the context can be a user’s query, the actions are recommended products, and the reward is the indicator of a purchase (Yue & Guestrin, 2011; Kveton et al., 2015; Combes et al., 2015; Zong et al., 2016; Li et al., 2016).

In many practical problems, the action space is large and cannot be explored naively. However, the mean rewards of actions are correlated due to some underlying structure. As a result, exploration of one action can teach the agent about other actions, depending on how correlated they are, and this increases statistical efficiency of exploration. As an example, in online shopping, many products are semantically similar, and their relationships can be captured by a graphical model: both a keyboard and monitor are computer accessories; and both computer accessories and home theatre systems are electronic devices. Another example is classification with a bandit feedback, where the labels are clustered. For instance, the car and truck are vehicles, while the monkey and tiger are animals. We experiment with this kind of problems in Section 6.2.

Unfortunately, the aforementioned structure is not easy to represent in traditional bandit algorithms (Auer et al., 2002; Chapelle & Li, 2012; Kawale et al., 2015; Sen et al., 2017). As an example, Thompson sampling (TS) (Thompson, 1933; Chapelle & Li, 2012; Agrawal & Goyal, 2012; Russo & Van Roy, 2014) is a natural and elegant strategy for exploration in bandit problems. However, in structured action spaces, TS would need to model correlations between all actions using their joint posterior. This could be computationally costly in practice, as observing the reward of a single action may require recomputing the joint posterior of all actions. Therefore, it is not immediately obvious how to design an efficient TS algorithm when the mean rewards of actions are correlated.

In this work, we study a structured bandit problem where the action space has a hierarchical structure. This structure is represented using a hierarchical Bayesian model (Lindley & Smith, 1972; Zhang & Yang, 2017), where each action is a leaf node in a tree. Each node is associated with a node parameter, which is drawn i.i.d. from a distribution parameterized by its parent’s parameter. This hierarchical structure not only models many practical bandit tasks, such as online shopping, but admits an efficient exploration because the...
joint posterior can be factorized and efficiently updated.

We make the following contributions. First, we formalize a hierarchical Bayesian model \( \mathcal{T} \) of our environment. Second, we propose a *Thompson sampling* algorithm HierTS. The main novelty in the design of HierTS is that the posterior factors along \( \mathcal{T} \), which permits exact *posterior sampling* and computationally-efficient updates. The posterior has a closed form in multi-armed and contextual linear bandits with Gaussian rewards. We derive Bayes regret bounds for HierTS that capture the structure of \( \mathcal{T} \) and the impact of priors. The bounds show increased statistical efficiency due to the hierarchy and can improve upon classical results in non-constant factors. We validate these theoretical results empirically and also apply HierTS to a real-world classification problem with label hierarchy.

2. Setting

We use the following notation. Random variables are capitalized. For any positive integer \( n \), we denote by \( [n] \) the set \( \{1, \ldots, n\} \). We let \( 1\{\cdot\} \) be the indicator function. The \( i \)-th entry of vector \( v \) is \( v_i \). If the vector is already indexed, such as \( v_j,i \), we write \( v_{j,i} \). We use \( \bar{O} \) for the big O notation up to logarithmic factors.

We consider a learning agent that interacts with a contextual bandit over \( n \) rounds (Li et al., 2010; Chu et al., 2011). In round \( t \in [n] \), the agent observes context \( X_t \in \mathcal{X} \subseteq \mathbb{R}^d \), takes an action \( A_t \) from an action set \( \mathcal{A} \) of size \( K \), and then observes a stochastic reward \( Y_t = r(X_t, A_t) + \varepsilon_t \), where \( r: \mathbb{R}^d \times \mathcal{A} \rightarrow \mathbb{R} \) is a reward function and \( \varepsilon_t \) is independent \( \sigma^2 \)-sub-Gaussian noise.

Our problem is structured. In particular, the action set \( \mathcal{A} \) progressively breaks into finer clusters of actions with similar rewards. This decomposition is represented by a tree \( \mathcal{T} \) (Figure 1) over nodes \( \mathcal{V} \subseteq \mathbb{N} \). Without loss of generality, the root has index \( 1 \). Each leaf of \( \mathcal{T} \) corresponds to an action \( a \in \mathcal{A} \) and we call it an action node. Each internal node of \( \mathcal{T} \) has at least two and at most \( b \) children, where \( b \) is the branching factor. The height of the tree is \( h \) and the height of node \( i \in \mathcal{V} \) is \( h_i \leq h \). The height of the leaves is 0 and the height of the root is \( h \). For any node \( i \in \mathcal{V} \), we denote its parent by \( \text{pa}(i) \) and its children by \( \text{ch}(i) \). An ancestor of node \( i \) is any node on a direct path from node \( i \) to the root, and node \( i \) is a descendant of any node on that path. With a slight abuse of notation, we use \( \mathcal{A} \subseteq \mathcal{V} \) to refer to both the action set and leaves of \( \mathcal{T} \), and sometimes index action nodes by \( a \) to stress their role.

The reward function is parameterized by model parameters \( \Theta = (\theta_i)_{i \in \mathcal{V}} \), where \( \theta_i \) is the parameter of node \( i \). The true model parameters are \( \Theta_* = (\theta_*; i \in \mathcal{V}) \) and we assume that they are generated as

\[
\theta_{*,1} \sim P_{0,1},
\theta_{*,i} \mid \theta_{*,\text{pa}(i)} \sim P_{0,i}(\cdot \mid \theta_{*,\text{pa}(i)}), \quad \forall i \in \mathcal{V} \setminus \{1\},
Y_t \mid X_t, \theta_{*,A_t} \sim P(\cdot \mid X_t; \theta_{*,A_t}), \quad \forall t \in [n].
\]

Here \( P_{0,1} \) is the prior distribution of the root node, called a hyper-prior; \( P_{0,i}(\cdot \mid \theta_{*,\text{pa}(i)}) \) is the conditional prior distribution of node \( i \), parameterized by the sampled value of its parent \( \theta_{*,\text{pa}(i)} \); and \( P(\cdot \mid x; \theta_{*,a}) \) is the reward distribution of action \( a \) in context \( x \). We use \( r(x, a; \Theta) \) to denote the mean reward of action \( a \) in context \( x \) under model parameters \( \Theta \) and define it as \( r(x, a; \Theta) = \mathbb{E}_{Y \sim P(\cdot \mid x, \theta_{*,a})}[Y] \). Thus \( r(x, a; \Theta) \) depends only on one parameter in \( \Theta \). The generative process in (1) relates any two node parameters to each other, through the lowest common ancestor. This induces complex correlations that can be used for efficient exploration. We discuss motivating examples for this setting in Section 1.

The goal is to minimize the \( n \)-round regret defined as

\[
\mathcal{R}(n; \Theta_*) = \mathbb{E} \left[ \sum_{t=1}^{n} r(X_t, A_{t,*}; \Theta_*) - r(X_t, A_t; \Theta_*) \right],
\]

where \( A_{t,*} = \arg \max_{a \in \mathcal{A}} r(X_t, a; \Theta_*) \) is the optimal action in round \( t \) given context \( X_t \). In this work, we assume that the parameters \( \Theta_* \) are also random. We define the \( n \)-round Bayes regret as \( \mathcal{BR}(n) = \mathbb{E} [\mathcal{R}(n; \Theta_*)] \), which takes an additional expectation over \( \Theta_* \). While weaker than the traditional frequentist regret \( \mathcal{R}(n; \Theta_*) \), the Bayes regret is a practical performance measure, when the average performance across multiple instances of model parameters is of interest (Russo & Van Roy, 2014; Hong et al., 2020).
Deep Hierarchy in Bandits

Algorithm 1 \texttt{HierTS}: Hierarchical Thompson sampling.

\begin{algorithm}
\caption{Hierarchical Thompson sampling.}

\textbf{Input:} Tree $T$ with height $h$, all priors $P_0$, in (1).

Initialize all posteriors $P_1$, $\leftarrow P_0$.

\textbf{for} $t = 1, \ldots, n$ \textbf{do}
  \textbf{for} $\ell = h - 1, \ldots, 0$ \textbf{do}
    \textbf{for} $i \in V_\ell$ \textbf{do}
      Sample $\theta_{t,i} \sim P_{t,i}$
    \textbf{end for}
  \textbf{end for}
  \textbf{end for}

$\Theta_t \leftarrow (\theta_{t,i})_{i \in V}$

Take action $A_t \leftarrow \arg\max_{a \in A} P(X_t; a; \Theta_t)$

Observe $Y_t \sim P(h \mid X_t; \theta_{*,A_t})$

\textbf{Compute new posteriors} $P_{t+1}$.

\end{algorithm}

3. Algorithm

Since our environment is a graphical model (Figure 1), we explore using Thompson sampling (TS) (Thompson, 1933; Chapelle & Li, 2012; Agrawal & Goyal, 2012; Russo & Van Roy, 2014). The main challenge in our algorithm design are \textit{latent variables}. Specifically, the rewards of action nodes are observed and permit direct learning of their parameters, $\theta_{*,a}$ for $a \in A$. In contrast, parameters $\theta_{*,i}$ of the internal nodes $i \in V \setminus A$ are only indirectly observed through their descendant action nodes, and thus are latent.

It is unclear if modeling of the latent variables is necessary. To discuss alternatives, we need to introduce some notation. Let $H_t = (X_t, A_t, Y_t)_{t \in [t-1]}$ be the history of all interactions of the agent until round $t$ and $\Theta_{*,A} = (\theta_{*,a})_{a \in A}$ be the true model parameters of the action nodes. Because the mean rewards of actions depend only on $\Theta_{*,A}$, we can use it as the natural solution to our problem. The computational efficiency of our approach and is also used in the regret analysis (Section 5).

Assuming that all posteriors $P_{t,i}$ can be computed efficiently, it is trivial to propose a hierarchical Thompson sampling algorithm for our problem. We call it \texttt{HierTS} and present its pseudo-code in Algorithm 1. In round $t$, \texttt{HierTS} works as follows. First, we sample the root parameter $\theta_{1,1}$. After that, we iterate over all nodes and sample node parameters whose parents are already sampled. Specifically, we define $V_\ell = \{i \in V : h_i = \ell\}$ as the subset of nodes at height $\ell$ and then sample $\theta_{t,i}$ for $i \in V_\ell$, starting from the children of the root at height $\ell = h - 1$ all the way to the leaves at height $\ell = 0$. In this design, $\Theta_t = (\theta_{t,i})_{i \in V}$ is a valid posterior sample, generated hierarchically. Finally, \texttt{HierTS} takes an optimistic action with respect to $\Theta_t$, observes $Y_t$, and then updates its posterior.

Note that \texttt{HierTS} samples parameters at all action nodes. It is possible to leverage the tree structure to prune sub-trees with actions that are unlikely to have high mean rewards. For example, Sen et al. (2021) propose beam search over a tree to only evaluate a subset of actions. Such computational improvements can be easily incorporated into \texttt{HierTS}. We view them as orthogonal to our main contribution, which is a statistically-efficient exploration using the tree structure.

3.2. Efficient Posterior Computation

The main technical novelty in \texttt{HierTS} is that the posteriors $P_{t,i}$ can be maintained efficiently. We show it as follows.

Fix any node $i$, its value $\theta$, and the value of its parent $\theta_p$. By Bayes rule, we have

$$P_{t,i}(\theta \mid \theta_p) \propto L_{t,i}(\theta) P_{0,i}(\theta \mid \theta_p),$$

where $L_{t,i}(\theta) = P(H_{t,i} \mid \theta_{*,i} = \theta)$ is the likelihood of observations $H_{t,i}$ whose ancestor is node $i$ with value $\theta$. Note that $L_{t,i}(\theta)$ can be further decomposed as

$$L_{t,i}(\theta) = \prod_{j \in ch(i)} \hat{L}_{t,j}(\theta),$$

where $\hat{L}_{t,j}(\theta) = P(H_{t,j} \mid \theta_{*,pa(j)} = \theta)$ is the likelihood of observations $H_{t,j}$ whose ancestor is child node $j$ and the value of its parent is $\theta$. This identity follows from two facts. First, $\theta_{*,j}$ are conditionally independent of each other given call a \textit{hyper-posterior}. For any other node $i$,

$$P_{t,i}(\theta \mid \theta_p) = P(\theta_{*,i} = \theta \mid \theta_{*,pa(i)} = \theta_p, H_{t,i})$$

is the posterior distribution of its parameter conditioned on $\theta_{*,pa(i)} = \theta_p$ in round $t$, where $H_{t,i}$ is a subset of interactions in $H_t$ where $A_t$ is a descendant of node $i$. In (2), $H_t$ could be replaced by $H_{t,i}$ because the posterior of $\theta_{*,i}$ is independent of the other observations given the value of the parent parameter $\theta_p$. This structure is critical to the computational efficiency of our approach and is also used in the regret analysis (Section 5).

3.1. Hierarchical Sampling

Our approach is based on hierarchical sampling (Andrieu et al., 2003; Doucet et al., 2001), where the model parameters $\Theta$ are sampled similarly to the generative process in (1). To explain it, we introduce the following notation. For the root node, $P_{t,1}(\theta) = P(\theta_{*,1} = \theta \mid H_t)$ denotes the posterior distribution of its parameter in round $t$, which we also
Algorithm 2 Updated statistics in round $t$. The dot notation means that the likelihoods are updated for all parameter values, which is possible in Gaussian models (Section 4).

**Initialize** $L_{t+1}.i \leftarrow L_t.i$.

$i \leftarrow A_t$

$L_{t+1.1}(\cdot) \leftarrow P(Y_t | X_t;i) L_{t.1}(\cdot)$

$L_{t+1.2}(\cdot) \leftarrow \int \theta L_{t+1.i}(\theta) P_{0,i}(\theta | \cdot) \, d\theta$

**repeat**

$i \leftarrow pa(i)$

$L_{t+1,i}(\cdot) \leftarrow \prod_{j \in ch(i)} L_{t+1.j}(\cdot)$

**if** $i > 1$

$L_{t+1,i}(\cdot) \leftarrow \int \theta L_{t+1.i}(\theta) P_{0,i}(\theta | \cdot) \, d\theta$

**end if**

**until** $i = 1$

Finally, each $L_{t,j}(\theta)$ can be computed as

$$
\hat{L}_{t,j}(\theta) = \int \theta L_{t,j}(\theta') P_{0,j}(\theta' | \theta) \, d\theta',
$$

(5)

where $L_{t,j}(\theta')$ is the likelihood of observations $H_{t,j}$ whose ancestor is node $j$ with value $\theta'$. Note that $L_{t,j}(\theta')$ can be further rewritten as in (4), which gives rise to our recursive computation of the posterior.

The pseudo-code for updating $L_t.i$ and $\hat{L}_t.i$ after round $t$ is shown in Algorithm 2. After this, (3) has to be recomputed for all nodes $i$ on the path from $A_t$ to the root. In general, (5) is hard to compute due to the integral over $\theta'$. However, in Gaussian graphical models (Section 4), this can be done in a closed form. In practice, (5) can be approximated for arbitrary distributions using approximate inference, either variational or MCMC (Doucet et al., 2001).

4. Gaussian Hierarchy

In this section, we instantiate the environment in (1) as a hierarchical Gaussian model (Koller & Friedman, 2009) and derive its posterior. The model is defined as

$$
\theta_{s,1} \sim \mathcal{N}(\mu_1, \Sigma_{0,1}),
$$

(6)

$$
\theta_{s,i} | \theta_{s,pa(i)} \sim \mathcal{N}(\theta_{s,pa(i)}, \Sigma_{0,i}), \quad \forall i \in \mathcal{V} \setminus \{1\},
$$

$$
Y_t | X_t, \theta_{s,A_1} \sim \mathcal{N}(X_t^T \theta_{s,A_1}, \sigma^2), \quad \forall t \in [n],
$$

where $\theta_{s,i} \in \mathbb{R}^d$ are the node parameters and $\Sigma_{0,i}$ are the covariance matrices that control the closeness of $\theta_{s,i}$ and $\theta_{s,pa(i)}$. The mean reward is defined as $r(x, a; \theta) = x^T \theta_a$.

The hierarchical structure is motivated by multi-label classification (Prabhu et al., 2018; Yu et al., 2020a), where $X_t$ is a feature vector, $A_t$ is its predicted label, and $Y_t$ indicates if the label is correct. We return to this application in Section 6.2. When $d = 1$ and $X_t = 1$, we recover a $K$-armed Gaussian bandit, where $\theta_{s,a}$ is the mean reward of action $a$. We assume that the agent knows the hyper-prior mean $\mu_1$, all covariances $\Sigma_{0,i}$, and reward noise $\sigma$. This is only used in the regret analysis, where we study exact posterior sampling. In our experiments (Section 6.2), we learn these quantities from data.

The special case of multi-armed bandits also shows computational gains over naive posterior sampling. Specifically, due to the dependencies in (1), $\theta_{s,A} | H_t$ is a multivariate Gaussian with $K$ dimensions. To sample from it, we can compute the root of the posterior covariance and multiply it by a $K$-dimensional standard normal vector, which takes $O(K^2)$ time. In contrast, sampling in HierTS takes $O(|\mathcal{V}|)$ time. When each internal node of $T$ has at least 2 children, which is without loss of generality, $|\mathcal{V}| \leq 2K$ and our computational gain is $O(K^2)$. Now we present closed-form posteriors for hierarchies of Gaussian and contextual linear models.

4.1. Multi-Armed Bandit

We start with a $K$-armed Gaussian bandit. In this setting, each node $i \in \mathcal{V}$ is associated with a single scalar parameter $\theta_{s,i} \in \mathbb{R}$, and its initial uncertainty is described by conditional prior variance $\Sigma_{0,i} = \sigma_{0,i}^2 \in \mathbb{R}$. The posteriors for this model are derived in Appendix A.1 and stated below.

For any node $i$, the posterior of $\theta_{s,i}$ given $\theta_{s,pa(i)} = \theta_p$ is $P_{t,i}(\theta | \theta_p) = \mathcal{N}(\hat{\theta}_{t,i}, \hat{\sigma}^2_{t,i})$. If node $i$ is an internal node,

$$
\hat{\sigma}^{-2} = \sigma_{0,i}^{-2} + \sum_{j \in ch(i)} \hat{\sigma}^{-2}_{t,j},
$$

(7)

$$
\hat{\theta}_{t,i} = \hat{\sigma}^2_{t,i} \left( \sigma_{0,i}^2 \hat{\theta}_p + \sum_{j \in ch(i)} \hat{\sigma}^{-2}_{t,j} \hat{\theta}_{t,j} \right).
$$

At the root ($i = 1$), we set $\theta_p = \mu_1$. The child parameters $\hat{\theta}_{t,j}$ and $\hat{\sigma}^2_{t,j}$ are computed recursively as follows. If node $j$ is an action node, then

$$
\hat{\sigma}^2_{t,j} = \hat{\sigma}^2_{0,j} + \frac{\sigma^2}{|S_{t,j}|},
$$

(8)

$$
\hat{\theta}_{t,j} = \frac{1}{|S_{t,j}|} \sum_{\ell \in S_{t,j}} Y_{\ell},
$$

where $S_{t,j} = \{ \ell < t : A_\ell = j \}$ are the rounds where action $j$ is taken before round $t$. If node $j$ is an internal node,

$$
\hat{\sigma}^2_{t,j} = \hat{\sigma}^2_{0,j} + M^{-1},
$$

$$
\hat{\theta}_{t,j} = M^{-1} \sum_{k \in ch(j)} \hat{\sigma}^{-2}_{t,k} \hat{\theta}_{t,k},
$$

(9)
Deep Hierarchy in Bandits

The recursive update in (9) can be derived from the observation that \( \tilde{L}_{t,j}(\theta) \propto \exp \left[ -\frac{1}{2} \tilde{\sigma}^{-2}_{t,j} \left( \theta - \hat{\theta}_{t,j} \right)^{2} \right] \) holds for any node \( j \) and the value of its parent \( \theta \). The closed-form of the posterior \( P_{t,i}(\theta | \theta_p) \) is a direct combination of this result and the derivations in Section 3.2. The update in (9) also has an intuitive interpretation. Although we get Gaussian observations at action nodes, as in (8), they propagate to higher nodes in the tree through (9). These nodes then act as noisy observations of their parents with mean \( \hat{\theta}_{t,j} \) and variance \( \tilde{\sigma}^{2}_{t,j} \). This allows us to overcome the problem of latent variables in our model. The posterior in (7) is just a function of higher-level observations in all children of node \( i \). To have closed forms of these quantities, we rely heavily on the properties of Gaussian random variables.

### 4.2. Contextual Linear Bandit

We now consider the general case in (6). This model can be viewed as a hierarchy of linear models (Yue & Guestrin, 2011; Abbasi-Yadkori et al., 2011) indexed by actions. The posteriors for this model are derived in Appendix A.2 and stated below.

For any node \( i \), the posterior of \( \theta_{s,i} \) given \( \theta_{s,pai}(i) = \theta_p \) is

\[
P_{t,i}(\theta | \theta_p) = N(\theta; \hat{\theta}_{t,i}, \tilde{\Sigma}_{t,i}).
\]

If node \( i \) is an internal node, then

\[
\hat{\Sigma}_{t,i}^{-1} = \hat{\Sigma}_{t,i}^{-1} + \sum_{j \in \text{ch}(i)} \hat{\Sigma}_{t,j}^{-1},
\]

\[
\hat{\theta}_{t,i} = \hat{\Sigma}_{t,i} \left( \sum_{j \in \text{ch}(i)} \hat{\Sigma}_{t,j}^{-1} \hat{\theta}_{t,j} \right).
\]

At the root \( (i = 1) \), we set \( \theta_p = \mu_1 \). The child parameters \( \hat{\theta}_{t,j} \) and \( \hat{\Sigma}_{t,j} \) are computed recursively as follows. If node \( j \) is an action node, then

\[
\hat{\Sigma}_{t,j} = \Sigma_{0,j} + G_{t,j}^{-1},
\]

\[
\hat{\theta}_{t,j} = \sigma^{-2} G_{t,j}^{-1} \sum_{\ell \in S_{t,j}} X_{\ell} Y_{\ell},
\]

where \( S_{t,j} = \{ \ell < t : A_{\ell} = j \} \) are the rounds where action \( j \) is taken before round \( t \) and \( G_{t,j} = \sigma^{-2} \sum_{\ell \in S_{t,j}} X_{\ell}^{T} X_{\ell} \) is the outer product of the corresponding feature vectors. If

node \( j \) is an internal node, then

\[
\hat{\Sigma}_{t,j} = \Sigma_{0,j} + M^{-1},
\]

\[
\hat{\theta}_{t,j} = M^{-1} \sum_{k \in \text{ch}(j)} \hat{\Sigma}_{t,k}^{-1} \hat{\theta}_{t,k},
\]

where \( M = \sum_{k \in \text{ch}(j)} \hat{\Sigma}_{t,k}^{-1} \). The new child parameters \( \hat{\theta}_{t,k} \) and \( \hat{\Sigma}_{t,k} \) are computed recursively, depending on whether \( k \) is an action node or not. Finally, if node \( i \) is action node, its posterior has a standard form of

\[
\hat{\Sigma}_{t,i}^{-1} = \Sigma_{0,i} + G_{t,i},
\]

\[
\hat{\theta}_{t,i} = \hat{\Sigma}_{t,i} \left( \sum_{j \in \text{ch}(i)} \hat{\Sigma}_{t,j}^{-1} \theta_p + \sigma^{-2} \sum_{\ell \in S_{t,i}} X_{\ell} Y_{\ell} \right).
\]

The recursive update in (12) can be derived from the observation that \( \tilde{L}_{t,j}(\theta) \propto \exp \left[ -\frac{1}{2} \tilde{\sigma}^{-2}_{t,j} \left( \theta - \hat{\theta}_{t,j} \right)^{2} \right] \) holds for any node \( j \) and the value of its parent \( \theta \). The second term in Lemma 1 is constant in \( \delta \). Since our model is a Gaussian graphical model, this posterior is a multivariate Gaussian (Koller & Friedman, 2009). This is because any conditioning or marginalization does not change the model class. This observation allows us to prove the following lemma.

**Lemma 1.** For any \( \delta > 0 \), the Bayes regret of HierTS is bounded as

\[
BR(n) \leq \sqrt{2nG(n)} \log(1/\delta) + \sqrt{2/\pi} \sigma_{\max} Kn\delta,
\]

where \( G(n) = E \left[ \sum_{t=1}^{n} \tilde{\sigma}^{2}_{t,A_{t}} \right] \) is a statistical complexity term, \( \tilde{\sigma}^{2}_{t,A_{t}} = \text{var} [ \theta(A_{t}) | H_{t} ] \) is the marginal posterior variance of the mean reward of action \( A_{t} \) in round \( t \), and \( \sigma_{\max} \) is the maximum marginal prior width at an action node.

The second term in Lemma 1 is constant in \( n \) for \( \delta = 1/n \). Therefore, we focus on the first \( O(\sqrt{n}) \) term. Also note that
For an internal node $i$

**Theorem 4.** The constant $\sigma_{0,i}$ in $\mathcal{G}(n)$ is none of the conditional posterior variances derived in Section 4.1. We show how to decompose it into these variances next.

To relate the marginal posterior variance, which bounds the expected regret, to conditional posterior variances, which represent our model uncertainty, we adopt the following update-centric notation. We denote the list of nodes from the root to the action node $A_i$ in round $t$ by $\psi_t$. The length of $\psi_t$ is $L_t$. To illustrate the notation, $\psi_t(1) = 1$ is the root, $\psi_t(L_t) = A_i$ is the action node, and $\psi_t(L_t - 1) = pa(A_i)$ is its parent. Figure 1 visualizes $\psi_t$. Now we are ready to relate the two quantities.

**Lemma 2.** In any round $t$, the marginal posterior variance in action node $A_i$ decomposes as

$$\bar{\sigma}_{t,A_i}^2 = \sum_{i=1}^{L_t} \left( \prod_{j=i+1}^{L_t} \frac{\hat{\sigma}_{t,\psi_t(j)}^4}{\sigma_{0,\psi_t(j)}^4} \right) \bar{\sigma}_{t,\psi_t(i)}^2.$$  

The last piece is a lower bound, which shows that each term in Lemma 2 can be bounded by the posterior update of the corresponding node $i$, representing our information gain.

**Lemma 3.** Fix any round $t$ and $i \in [L_t]$. Then

$$\bar{\sigma}_{t,\psi_t(i)}^2 - \bar{\sigma}_{t,\psi_t(i)}^2 \geq c^{-2} \sum_{j=i+1}^{L_t} \left( \prod_{j=i+1}^{L_t} \frac{\hat{\sigma}_{t,\psi_t(j)}^4}{\sigma_{0,\psi_t(j)}^4} \right),$$

where $c = 1 + \sigma_{0,\max}^2 / \sigma^2$ and $\sigma_{0,\max}^2 = \max_{i \in V} \sigma_{0,i}^2$ is the maximum prior variance. For $\sigma \geq \sigma_{0,\max}$, we have $c = 2$.

The constant $c$ in Lemma 3 is small when the reward noise is higher than all prior widths in $T$. This property can be always attained by initial forced exploration of all actions.

### 5.2. Regret Bound

Now we are ready to present our main result. Recall that $h$ is the height of $T$, $h_i$ is the height of node $i$, and that the action nodes have height 0 (Section 2).

**Theorem 4.** For any $\delta > 0$, the Bayes regret of HiérTS is bounded as

$$BR(n) \leq \sqrt{2nG(n)\log(1/b)} + \sqrt{2/\pi \sigma_{\max}^2 Kn\delta},$$

where $G(n) = \sum_{i \in V} c_i w_i$ and $c$ is a constant defined in Lemma 3. For an action node $i$, $h_i = 0$ and

$$w_i = \frac{\sigma_{0,i}^2}{\log \left( 1 + \frac{\sigma_{0,i}^2 n}{\sigma^2} \right)} \log \left( 1 + \frac{\sigma_{0,i}^2 n}{\sigma^2} \right).$$

For an internal node $i$, $h_i > 0$ and

$$w_i = \frac{\sigma_{0,i}^2}{\log \left( 1 + \frac{\sigma_{0,i}^2}{\sigma^2} \right)} \log \left( 1 + \frac{\sigma_{0,i}^2}{\sigma^2} \right) \sum_{j \in ch(i)} \sigma_{0,j}^2.$$  

When $\delta = 1/n$, the above bound is $\tilde{O}(\sqrt{n|V|})$, where $n$ is the horizon and $|V|$ is the number of nodes, which is also the number of learned parameters. The dependence on the horizon $n$ is standard. As $w_i = \tilde{O}(\sigma_{0,i}^2)$, the contribution of each node $i$ to the regret is proportional to its prior width. Thus the regret decreases when the model is more certain. One notable dependence in $\mathcal{G}(n)$ is exponential scaling with height $h_i$. This is not problematic, as the number of nodes with high $h_i$ is exponentially smaller than those with lower $h_i$ (Section 5.3).

Theorem 4 also recovers a well-known Bayes regret bound for $K$-armed bandits (Russo & Van Roy, 2014). The reason is that a $K$-armed bandit can be viewed as a tree with height $h = 1$, where the root parameter $\theta_1$ is the prior mean of the actions. Because $\theta_1$ is certain, $w_1 = 0$ and $\sum_{i \in V} c_i w_i = \sum_{i=2}^{K+1} w_i = O(K)$.

### 5.3. Lower Regret Due to Hierarchy

Now we give examples of how the hierarchy can help with reducing regret. To simplify the discussion, we ignore logarithmic factors in the definitions of $w_i$ in Theorem 4. We assume that $T$ is a balanced $b$-ary tree with height $h$; with $K = b^h$ action nodes and $b^{h-\ell}$ nodes at height $\ell$. Our discussion is under the assumption that $c = 2$, as derived in Lemma 3. More gains are possible when $c < 2$.

We compare the regret of HiérTS to classical Thompson sampling (TS), which maintains an independent posterior of $\theta_{*,a}$ for each action $a \in A$. To have a fair comparison, we set the marginal prior variances of all actions in TS as in HiérTS. Specifically, let $\psi_a$ be the path in $T$ from action node $a$ to the root. Then the marginal prior of action $a$ is $N(\mu_1, \sigma_{0,a}^2)$, where $\sigma_{0,a}^2 = \sum_{i \in \psi_a} \sigma_{0,i}^2$ and $\mu_1$ denotes the hyper-prior mean in (6). The regret of this algorithm can be bounded using Theorem 4, with the only difference that the complexity term becomes $G_{TS}(n) \approx \sum_{a \in A} \sigma_{0,a}^2$.

**Problem 1.** We start with a problem where all prior variances are identical, $\sigma_{0,i}^2 = 1$ for any $i \in V$. In this case, all prior variances in TS are $\sigma_{0,a}^2 = h + 1$ and its complexity term is $G_{TS}(n) = (h + 1)b^h$. In HiérTS, we aggregate the nodes by height and get

$$G(n) = \sum_{\ell=0}^{h} b^{h-\ell} c^\ell = b^h \sum_{\ell=0}^{h} (2/b)\ell \leq \frac{1}{1-2/b} b^h.$$  

Thus HiérTS improves $G(n)$ by $\Omega(h)$ when $b > 2$. Since $h = \log_b b = \log_b K$, we get $G_{TS}(n)/G(n) \approx \log_b K$, and HiérTS reduces the Bayes regret by a multiplicative factor $\sqrt{\log_b K}$. This argument can be adjusted for $b = 2$ to get a comparable regret to TS.

**Problem 2.** Now we consider a problem where the conditional prior variances in $T$ double with height, $\sigma_{0,i}^2 = 2^{h_i}$,
where \( h_i \) is the height of node \( i \). This setting is motivated in Section 1. We expect higher statistical gains because the uncertainty of highly-uncertain nodes at higher levels of \( T \) is reduced jointly by all actions. In this problem, all prior variances in \( \mathcal{T} \) are \( \sigma^2_{0,a} = 2^{h+1} \) and its complexity term is \( G_{\mathcal{T}}(n) = 2^{h+1}b^h \). In comparison, \( \text{HierTS} \) yields

\[
G(n) = \sum_{\ell=0}^{h} 2^{h-\ell}c^{2\ell} \leq \sum_{\ell=0}^{h} 2^{h-\ell}4^{\ell} \leq \frac{1}{1-4/b}b^h,
\]

where the last step is analogous to Problem 1. So \( \text{HierTS} \) improves \( G(n) \) by \( \Omega(2^{h+1}) \) if \( b > 4 \). Since

\[
2^{h+1} = 2 \cdot 2^{\log_b b} = 2 \cdot 2^{\log_b b / \log_b 2} = 2K \frac{1}{\log_2 b},
\]

we get \( G_{\mathcal{T}}(n)/G(n) \approx K^{1/\log_2 b} \), and \( \text{HierTS} \) reduces the Bayes regret by a multiplicative factor \( \sqrt{K^{1/\log_2 b}} \). For \( b = 5 \), this factor would be close to \( K^{1/2} \), Therefore, the regret is reduced by a polynomial factor in \( K \).

### 5.4. Contextual Regret Bound

This section presents a contextual generalization of the regret bound in Theorem 4. We make two assumptions in its derivation. First, the length of the context vector is bounded, and we assume that \( \|X_t\|_2 \leq 1 \) without loss of generality. Second, \( \Sigma_{0,i} = \tilde{\sigma}^2_{0,i} I_d \). This latter assumption allows us to utilize \( \lambda_1(\Sigma_{0,i}) = \lambda_d(\Sigma_{0,i}) = \sigma^2_{0,i} \), which is required by our matrix generalizations of Lemmas 2 and 3. Our regret bound is stated below.

**Theorem 5.** For any \( \delta > 0 \), the Bayes regret of \( \text{HierTS} \) is bounded as

\[
BR(n) \leq \sqrt{2dnG(n) \log(1/\delta)} + \sqrt{2/\pi \sigma_{\max} Kn\delta},
\]

where \( G(n) = \sum_{v \in V} c^{h_v}w_v \), and \( c \) is a constant defined in Lemma 3. For an action node \( i \), \( h_i = 0 \) and

\[
w_i = \frac{\sigma^2_{0,i}}{\log (1 + \sigma^2_{0,i})} \log \left( 1 + \sigma^2_{0,i,n} \right) / \sigma^2_{0,i}.
\]

For an internal node \( i \), \( h_i > 0 \) and

\[
w_i = \frac{\sigma^2_{0,i}}{\log (1 + \sigma^2_{0,i})} \log \left( 1 + \sigma^2_{0,i} \sum_{j \in ch(i)} \sigma_{0,j} \right).
\]

The constants are similar to Theorem 4. The main difference is the extra factor of \( \sqrt{d} \), because each node is now associated with a learned \( d \)-dimensional parameter.

The key insight in the proof is that the context in round \( t \), \( X_t \), is known and fixed. Because all conditional posteriors are Gaussian, posterior sampling in contextual \( \text{HierTS} \) can be implemented using a tree with scalar nodes, where the conditional posterior of node \( i \) is \( \mathcal{N}(X_i^\top \bar{\theta}_{i,t}, X_i^\top \Sigma_{i,t} X_i) \). As a result, the non-contextual analysis in Theorem 4 can be easily generalized. Lemma 1 changes in that the history \( H_t \) includes context \( X_t \). Moreover, the marginal posterior variance of the mean reward of action \( A_i \), \( \tilde{\sigma}^2_{i,t,A_i} \), turns into that in the direction of context \( X_t \).

The main technical challenges in the proof are generalizing the variance decomposition in Lemma 3 to covariances (Lemma 9), and extending the posterior update lower bound in Lemma 3 to matrices (Lemma 9). The rest of the proof follows the same outline as that of Theorem 4.

### 6. Experiments

We compare \( \text{HierTS} \) to three baselines that either neglect or partially use the tree \( T \). The first baseline is Thompson sampling (TS), which treats each action independently and is introduced in Section 5.3. The second baseline only uses a 2-level hierarchy, namely the root and action nodes, and we call it \( \text{FlatTS} \). Its hyper-prior for the root \( P_{0,1} \) is the same as in \( \text{HierTS} \). For any action \( a \in A \), the conditional prior is \( P_{0,a}(\cdot | \theta_{s,a}) = \mathcal{N}(\cdot | \theta_{s,a}, \tilde{\sigma}^2_{a} - \sigma^2_{0,a}) \), where \( \tilde{\sigma}^2_{a} \) is the marginal prior variance in TS. This baseline mimics existing algorithms for 2-level Gaussian hierarchies with a common root (Kveton et al., 2021; Basu et al., 2021; Hong et al., 2022), and is similar to structured bandits where the actions share a latent parameter (Gupta et al., 2018).

The final baseline is the zooming algorithm of Kleinberg et al. (2008; 2013), which we call Zooming. This is a UCB algorithm where actions are embedded in a metric space, and can be viewed as a standard approach to handling large structured action spaces in bandits (Bubeck et al., 2008; Kleinberg et al., 2013). At a high level, Zooming maintains a set of “active” actions, which are sufficiently apart from each other given the history, and only explores these. The key step in implementing Zooming is constructing a metric space where actions with similar mean rewards are close. We design it as follows. We compute the graph Laplacian of \( T \), represented as an undirected graph, and extract its \( d \) eigenvectors \( v_{i,j} \) with the smallest eigenvalues. Let \( v_{i,j} \) be the \( j \)-th entry of \( v_i \). Then, for any action node \( a \in A \), its embedding in the metric space is \( (v_{1,a}, \ldots, v_{d,a}) \). This is a standard approach to deriving a metric space in spectral clustering (Yan et al., 2009).

#### 6.1. Synthetic Experiments

Our first experiments are on a synthetic Gaussian bandit, where we validate theoretical findings from Section 5.3. We experiment with both problems in Section 5.3, which are \( b \)-ary trees with height \( h \) and \( K = b^h \) actions. In Problem 1, the prior variances are constant. In Problem 2, the prior
Deep Hierarchy in Bandits

Figure 2. Regret of HierTS on synthetic bandit problems with varying branching factor $b$.

variances double with height. In both problems, the mean of the hyper-prior is $\mu_1 = 0$, and the reward of action $a$ is $\theta_{s,a}$ with variance $\sigma^2 = 1$.

We start with Problem 2, where the tree height is $h = 2$ and we vary the branching factor $b$. All algorithms are run for $n = 500$ rounds and evaluated by the Bayes regret on 100 independent samples of $\Theta_s$. Its mean estimate and standard error are reported in Figure 2. For all values of $b$, HierTS outperforms all baselines. Zooming is not competitive with the other baselines when the branching factor $b$ is low. The poor performance of Zooming is because of its generality. Specifically, since it can be applied to any smooth reward function, its regret in $d$ dimensions scales as $\tilde{O}(n^1 + \sqrt{d})$.

In Figure 2, we only show the results for Zooming where $d = K$. We experimented with $d < K$ and observed linear regret, because some action nodes with a common parent had identical features.

In the next experiment, we consider both Problems 1 and 2. The branching factor is fixed at $b = 2$ and we vary the tree height $h$. All algorithms are run for $n = 500$ rounds on 100 independent samples of $\Theta_s$. The reduction in the Bayes regret of HierTS and FlatTS is measured as the ratio of the TS regret over the regret in question. In Figure 3a, we plot the ratios for Problem 1. Section 5.3 suggests that the regret decreases by $\Omega(\sqrt{h})$. Our plot confirms this. In Figure 3b, we plot the ratios for Problem 2. Section 5.3 suggests that the regret decreases by $\Omega(2^{h/2})$. We confirm this trend.

6.2. Multi-Label Image Classification

The last experiment is on a multi-label image classification problem with linear rewards. All compared algorithms are implemented analogously to Section 6.1, except for variances being replaced with covariances. We use the CIFAR-100 dataset (Krizhevsky, 2009), with 60000 images of size $32 \times 32$. There are 50000 training and 10000 test images. Each image belongs to one of 100 classes (labels) and 20 super-classes, each consisting of 5 classes. Each image is represented by a $d = 10$ dimensional feature vector, which we obtain by downsampling a 100-dimensional feature vector. That feature vector is an embedding computed by an EfficientNet-L2 network applied to the image (Tan & Le, 2019; Xie et al., 2020; Foret et al., 2021). The network is a convolutional neural network pretrained on both ImageNet (Russakovsky et al., 2015) and unlabeled JFT-300M (Sun et al., 2017), and fine-tuned on the CIFAR-100 training set.

We randomly select 5 super-classes and their corresponding $K = 25$ classes become actions. The test and training sets are restricted to these classes. Our bandit problem is defined as follows. For each action $a$, $\theta_{s,a}$ is the mean feature vector of test images in class $a$. In round $t$, the context $X_t$ is the feature vector of a random image from the test set and the reward of action $A_t$ is $Y_t \sim N(X_t^T \theta_{s,A_t}, 0.5^2)$. Therefore, the mean reward is maximized whenever the true class is chosen. Finally, we build a 3-level hierarchy $\mathcal{T}$ as follows. The hyper-prior of the root is a multivariate Gaussian fitted to all training images, $P_{0,1} = N(\mu_1, \Sigma_{0,1})$. The nodes at height 1 are the 5 super-classes and their conditional priors are $P_{0,i}(\cdot \mid \theta_{s,1}) = N(\cdot \mid \mu_{i,1}, \Sigma_{0,i})$, where $\Sigma_{0,i}$ is fitted to the training images of super-class $a$. Finally, the nodes at height 0 correspond to actions and their conditional priors are $P_{0,a}(\cdot \mid \theta_{s,pa(a)}) = N(\cdot \mid \mu_{a,pa(a)}, \Sigma_{0,a})$, where $\Sigma_{0,a}$ is fitted to the training images of class $a$.

We report the mean and standard error of the regret over 10 runs in Figure 4. We observe again that HierTS performs better than all baselines. We do not consider Zooming due to its poor performance earlier (Figure 2). Note that the true model parameters of $\mathcal{T}$, namely $\mu_1$ and $\Sigma_{0,i}$, are unknown; and we estimate them from training images. Therefore, this experiment shows that even when we relax the assumption that they are known, it is beneficial to estimate them, and use the structure of $\mathcal{T}$.

7. Related Work

Thompson sampling algorithms have been widely applied to contextual bandits because of their computational efficiency and strong empirical performance (Chu et al., 2011; Chapelle & Li, 2012; Abbasi-Yadkori et al., 2011). Russo & Van Roy (2014) proved first Bayes regret bounds for TS. Our proposed algorithm HierTS extends TS to tree hierarchies. TS with a 2-level hierarchy over tasks was applied
and analyzed in both meta-learning and multi-task learning (Kveto et al., 2021; Basu et al., 2021; Wan et al., 2021; Hong et al., 2022). The main difference in our work is that we move from 2-level hierarchies to an arbitrary depth, and develop both algorithmic and theory foundations for this setting. Our analysis extends the variance decompositions proposed in Hong et al. (2022) to trees. Alternatively, information theory could be used to derive Bayes regret bounds (Russo & Van Roy, 2016; Lu & Van Roy, 2019; Basu et al., 2021), but we are unaware of any for trees.

Our problem has a structured action space. One structure studied by prior works is a shared latent parameter among all actions (Tirinzoni et al., 2020; Lattimore & Munos, 2014; Gupta et al., 2018). This can be viewed as an instance of our setting with a 2-level hierarchy. In latent bandits, the parameter is a discrete variable (Maillard & Mannor, 2014; Hong et al., 2020). Recent works also applied approximate TS to more complex structures (Gopalan et al., 2014; Yu et al., 2020b). Such algorithms are general, but can only be analyzed in limited settings under strong assumptions. We consider a special tree structure, where we can derive and analyze an exact algorithm. Another recent work that applies trees to bandits is Majzoubi et al. (2020), who proposed a general reduction-based algorithm for contextual bandits with a continuous action space, where the policies are represented by a tree.

The most related work is Sen et al. (2021), who also studied contextual bandits with a tree hierarchy over actions. Both of our works rely on a hierarchy of regressors, motivated by multi-label classification (Prabhu et al., 2018; Yu et al., 2020a). However, the works differ in several key aspects. First, we study a stochastic variant of the problem, where the tree nodes are associated with prior distributions rather than fixed centers and radii, as in Sen et al. (2021). In fact, Sen et al. (2021) did not model the statistical uncertainty at all. Second, we propose a TS algorithm using novel recursive derivations of the posterior. Sen et al. (2021) proposed a greedy strategy and beam search to avoid evaluation of all actions. In Section 3, we discuss how similar improvements can be incorporated in HierTS. Finally, our Bayesian analysis reveals structural properties that imply low regret. The low regret in Sen et al. (2021) is attained by making an assumption on the regression oracle, and can grow linearly when the oracle is imperfect.

8. Conclusions

In many practical problems, the action space is large and a good generalization over actions is not obvious. Motivated by this, we study a contextual bandit problem with a deep hierarchy over actions. To solve the problem, we propose hierarchical Thompson sampling (HierTS), which can be implemented exactly and efficiently in Gaussian models. We prove Bayes regret bounds for HierTS that quantify its increased statistical efficiency over vanilla TS and validate this experimentally. We also apply HierTS to a challenging classification problem with label hierarchy.

Our work is a major step towards bandit algorithms with rich graphical models. Its limitations can be addressed by future works. For instance, a frequentist regret analysis is possible and would only differ in Lemma 1. The rest of the proof, which captures the structure of our problem, is a worst-case argument. Second, we believe that our method can be generalized beyond Gaussian trees. As discussed in Section 3.2, exact posterior sampling is challenging under the constraint of computational efficiency; but many tractable approximations exist. When exact sampling is possible, we believe that our proofs can be extended to general exponential-family distributions. Another direction for future work is an extension to directed acyclic graphs (DAGs). The nodes in DAGs can be ordered, and therefore similar recursions to Sections 4 and 5 can be established.

References


Agrawal, S. and Goyal, N. Analysis of Thompson sampling
Deep Hierarchy in Bandits


A. Posterior Derivations

This section contains our posterior derivations. We adopt the convention that $\Lambda = \Sigma^{-1}$, where $\Lambda$ is the precision matrix for covariance $\Sigma$.

A.1. Multi-Armed Bandit Posterior

The proof is by induction. We start with the inductive step.

**Lemma 6.** Fix an internal node $j$. Let $\mathcal{pa}(j) = i$ and $C = \text{ch}(j)$. Let

$$
\mathbb{P}(H_{i,j}, \theta_{*,j} = \theta | \theta_{*,i} = \theta_i) \propto \exp \left[ -\frac{1}{2} \left( \sigma^{-2}_0 (\theta - \theta_i)^2 + \sum_{k \in C} \sigma^{-2}_k (\theta - \theta_k)^2 \right) \right],
$$

where $\theta_k, \sigma_k$ are the parameters of $k \in C$. Then

$$
\mathbb{P}(H_{i,j} | \theta_{*,i} = \theta_i) \propto \exp \left[ -\frac{1}{2} \tilde{\sigma}^{-2}_j (\theta_i - \tilde{\theta}_j)^2 \right]
$$

for $\tilde{\sigma}^2_j = \frac{\sigma^2_0}{s} + \left( \sum_{k \in C} \sigma^{-2}_k \right)^{-1}$ and $\tilde{\theta}_j = \left( \sum_{k \in C} \sigma^{-2}_k \right)^{-1} \sum_{k \in C} \sigma^{-2}_k \theta_k$.

**Proof.** Let $s = \sigma^{-2}_0 + \sum_{k \in C} \sigma^{-2}_k$ and $v = \sigma^{-2}_0 \theta_i + \sum_{k \in C} \sigma^{-2}_k \theta_k$. We start with completing the square of $\theta$,

$$
\log \mathbb{P}(H_{i,j}, \theta_{*,j} = \theta | \theta_{*,i} = \theta_i) \propto \sigma^{-2}_0 (\theta - \theta_i)^2 + \sum_{k \in C} \sigma^{-2}_k (\theta - \theta_k)^2
$$

$$
\propto s \theta^2 - 2\theta \left( \sigma^{-2}_0 \theta_i + \sum_{k \in C} \sigma^{-2}_k \theta_k \right) + \sigma^{-2}_0 \theta_i^2
$$

$$
= s(\theta^2 - 2\theta \theta_i s^{-1} v + s^{-2} v^2) + \sigma^{-2}_0 \theta_i^2 - s^{-1} v^2
$$

$$
= s(\theta - s^{-1} v)^2 + \sigma^{-2}_0 \theta_i^2 - s^{-1} v^2.
$$

In the second step, we omit constants in $\theta$ and $\theta_i$. Since we got a quadratic form in $\theta$, we know that

$$
\int_\theta \mathbb{P}(H_{i,j}, \theta_{*,j} = \theta | \theta_{*,i} = \theta_i) \, d\theta \propto \exp \left[ -\frac{1}{2} \left( \sigma^{-2}_0 \theta_i^2 - s^{-1} v^2 \right) \right].
$$

Let $\hat{s} = \sigma^{-2}_0 - \sigma^{-4}_0 s^{-1}$. Now we complete the square of $\theta_i$,

$$
\sigma^{-2}_0 \theta_i^2 - s^{-1} v^2 = \sigma^{-2}_0 \theta_i^2 - s^{-1} \left( \sigma^{-2}_0 \theta_i + \sum_{k \in C} \sigma^{-2}_k \theta_k \right)^2
$$

$$
= \sigma^{-2}_0 \theta_i^2 - \sigma^{-2}_0 \left( \theta_i + \sum_{k \in C} \sigma^{-2}_k \theta_k \right)^2
$$

$$
\propto \hat{s} \left( \theta_i^2 - 2\theta_i \hat{s}^{-1} \sigma^{-2}_0 s^{-1} \sum_{k \in C} \sigma^{-2}_k \theta_k \right) \propto \hat{s} \left( \theta_i - \hat{s}^{-1} \sigma^{-2}_0 s^{-1} \sum_{k \in C} \sigma^{-2}_k \theta_k \right)^2.
$$

In the last two steps, we omit constants in $\theta_i$. Finally, note that

$$
\hat{s} = \frac{\sigma^{-2}_0 (s - \sigma^{-2}_0)}{s} = \left( \sigma^{-2}_0 + (s - \sigma^{-2}_0)^{-1} \right)^{-1},
$$

$$
\hat{s}^{-1} \sigma^{-2}_0 s^{-1} = \frac{s}{\sigma^{-2}_0 (s - \sigma^{-2}_0)} \sigma^{-2}_0 s^{-1} = (s - \sigma^{-2}_0)^{-1}.
$$

This completes the proof, for $\tilde{\theta}_j = (s - \sigma^{-2}_0)^{-1} \sum_{k \in C} \sigma^{-2}_k \theta_k$ and $\tilde{\sigma}^2_j = \sigma^2_0 + (s - \sigma^{-2}_0)^{-1}$.

At an action node $j$, (13) holds for $\sigma_k = \sigma$ and $\theta_k = Y_k$, where $Y_k$ is an observation $k$ of node $j$ and $\sigma$ is observation noise. This is the base case of our proof by induction.
A.2. Linear Bandit Posterior

The proof is by induction. We start with the inductive step.

**Lemma 7.** Fix an internal node \( j \). Let \( pa(j) = i \) and \( C = ch(j) \). Let

\[
\mathbb{P}(H_{t,j}, \theta_{*,j} = \theta | \theta_{*,i} = \theta_i) \propto \exp \left[ -\frac{1}{2} \left( (\theta - \theta_i)^\top \Lambda_0 (\theta - \theta_i) + \sum_{k \in C} (\theta - \theta_k)^\top \Lambda_k (\theta - \theta_k) \right) \right],
\]

where \( \theta_k, \Lambda_k \) are the parameters of \( k \in C \). Then

\[
\mathbb{P}(H_{t,j} | \theta_{*,i} = \theta_i) \propto \exp \left[ -\frac{1}{2} (\theta_i - \tilde{\theta}_j)^\top \tilde{\Lambda}_j (\theta_i - \tilde{\theta}_j) \right]
\]

for \( \tilde{\Lambda}_j^{-1} = \Lambda_0^{-1} + (\sum_{k \in C} \Lambda_k)^{-1} \) and \( \tilde{\theta}_j = (\sum_{k \in C} \Lambda_k)^{-1} \sum_{k \in C} \Lambda_k \mu_k \).

**Proof.** Let \( S = \Lambda_0 + \sum_{k \in C} \Lambda_k \) and \( V = \Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k \). We start with completing the square of \( \theta \),

\[
\log \mathbb{P}(H_{t,j}, \theta_{*,j} = \theta | \theta_{*,i} = \theta_i) \propto (\theta - \theta_i)^\top \Lambda_0 (\theta - \theta_i) + \sum_{k \in C} (\theta - \theta_k)^\top \Lambda_k (\theta - \theta_k)
\]

\[
\propto \theta^\top S \theta - 2 \theta^\top \left( \Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k \right) + \theta_i^\top \Lambda_0 \theta_i
\]

\[
= \theta^\top S (\theta - S^{-1} V) + \theta_i^\top \Lambda_0 \theta_i
\]

\[
= (\theta - S^{-1} V)^\top S (\theta - S^{-1} V) + \theta_i^\top \Lambda_0 \theta_i - V^\top S^{-1} V.
\]

In the second step, we omit constants in \( \theta \) and \( \theta_i \). Since we got a quadratic form in \( \theta \), we know that

\[
\int_{\theta} \mathbb{P}(H_{t,j}, \theta_{*,j} = \theta | \theta_{*,i} = \theta_i) \, d\theta \propto \exp \left[ -\frac{1}{2} (\theta_i^\top \Lambda_0 \theta_i - V^\top S^{-1} V) \right].
\]

Let \( \tilde{S} = \Lambda_0 - \Lambda_0 S^{-1} \Lambda_0 \). Now we complete the square of \( \theta_i \),

\[
\theta_i^\top \Lambda_0 \theta_i - V^\top S^{-1} V = \theta_i^\top \Lambda_0 \theta_i - \left( \Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k \right)^\top S^{-1} \left( \Lambda_0 \theta_i + \sum_{k \in C} \Lambda_k \theta_k \right)
\]

\[
\propto \theta_i^\top \tilde{S} \left( \theta_i - 2 \tilde{S}^{-1} \Lambda_0 S^{-1} \sum_{k \in C} \Lambda_k \theta_k \right)
\]

\[
\propto \left( \theta_i - \tilde{S}^{-1} \Lambda_0 S^{-1} \sum_{k \in C} \Lambda_k \theta_k \right)^\top \tilde{S} \left( \theta_i - \tilde{S}^{-1} \Lambda_0 S^{-1} \sum_{k \in C} \Lambda_k \theta_k \right).
\]

In the last two steps, we omit constants in \( \theta_i \). Finally, by the Woodbury matrix identity, we have

\[
\tilde{S} = \Lambda_0 - \Lambda_0 S^{-1} \Lambda_0 = (\Lambda_0^{-1} + (S - \Lambda_0)^{-1})^{-1},
\]

\[
\tilde{S}^{-1} \Lambda_0 S^{-1} = (\Lambda_0 - \Lambda_0 S^{-1} \Lambda_0)^{-1} \Lambda_0 S^{-1} = (S - \Lambda_0)^{-1}.
\]

This completes the proof, for \( \tilde{\theta}_j = (S - \Lambda_0)^{-1} \sum_{k \in C} \Lambda_k \mu_k \) and \( \tilde{\Lambda}_j = (\Lambda_0^{-1} + (S - \Lambda_0)^{-1})^{-1} \).

For any node \( j \), note that (14) can be written as

\[
\mathbb{P}(H_{t,j}, \theta_{*,j} = \theta | \theta_{*,i} = \theta_i) \propto \exp \left[ -\frac{1}{2} \left( (\theta - \theta_i)^\top \Lambda_0 (\theta - \theta_i) + \sum_{k \in C} \theta^\top \Lambda_k \theta - 2 \sum_{k \in C} \theta^\top \Lambda_k \theta_k \right) \right],
\]

when constants in \( \theta \) and \( \theta_i \) are omitted. Then, at an action node, \( \Lambda_k = \sigma^{-2} X_k^\top X_k \) and \( \Lambda_k \theta_k = \sigma^{-2} X_k Y_k \), where \( Y_k \) is an observation \( k \) of node \( j \) at feature vector \( X_k \) and \( \sigma \) is observation noise. This is the base case of our proof by induction.
B. Regret Bound Proofs

This section contains proofs of our regret bound and supporting lemmas.

B.1. Proof of Lemma 1

Fix round $t$. Let $\mathbb{P}(\Theta \mid H_t) = \mathcal{N}(\Theta; \bar{\Theta}_t, \bar{\Sigma}_t)$ be the joint posterior distribution of all action node parameters $\Theta \in \mathbb{R}^K$, with mean $\bar{\Theta}_t \in \mathbb{R}^K$ and covariance $\bar{\Sigma}_t \in \mathbb{R}^{K \times K}$. Let $A_t \in \{0, 1\}^K$ and $A_* \in \{0, 1\}^K$ be indicator vectors of the taken action in round $t$ and the optimal action, respectively. Each action is associated with one leaf node.

Since $\bar{\Theta}_t$ is deterministic given $H_t$, and $A_*$ and $A_t$ are i.i.d. given $H_t$, we have

$$\mathbb{E} \left[ A_*^\top (\Theta_* - \bar{\Theta}_t) \right] = \mathbb{E} \left[ A_*^\top (\Theta_* - \Theta_t) \mid H_t \right] + \mathbb{E} \left[ A_*^\top (\Theta_t - \bar{\Theta}_t) \mid H_t \right].$$

Moreover, $\Theta_* - \bar{\Theta}_t$ is a zero-mean random vector independent of $A_t$, and thus $\mathbb{E} \left[ A_*^\top (\Theta_* - \Theta_t) \mid H_t \right] = 0$. So we only need to bound the first term above. Let

$$E_t = \left\{ \forall a \in A : |a^\top (\Theta_* - \bar{\Theta}_t)| \leq \sqrt{2 \log(1/\delta)} \|a\|_{\bar{\Sigma}_t} \right\}$$

be the event that all high-probability confidence intervals hold. Fix history $H_t$. Then by the Cauchy-Schwarz inequality,

$$\mathbb{E} \left[ A_*^\top (\Theta_* - \bar{\Theta}_t) \mid H_t \right] \leq \sqrt{2 \log(1/\delta)} \mathbb{E} \left[ \|A_*\|_{\bar{\Sigma}_t} \mid H_t \right] + \mathbb{E} \left[ A_*^\top (\Theta_* - \bar{\Theta}_t) \mid E_t \right] \mathbb{P}(E_t \mid H_t).$$

Now note that for any action $a$, $a^\top (\Theta_* - \bar{\Theta}_t)/\|a\|_{\bar{\Sigma}_t}$ is a standard normal variable. It follows that

$$\mathbb{E} \left[ A_*^\top (\Theta_* - \bar{\Theta}_t) \mid H_t \right] \leq \sqrt{2 \log(1/\delta)} \mathbb{E} \left[ \|A_*\|_{\bar{\Sigma}_t} \mid H_t \right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K \delta,$$

where we use that $E_t$ implies $\|\Sigma_t^{\frac{1}{2}} (\Theta_* - \bar{\Theta}_t)\|_{\infty} \geq \sqrt{2 \log(1/\delta)}$. Now we combine all inequalities and have

$$\mathbb{E} \left[ A_*^\top (\Theta_* - \bar{\Theta}_t) \mid H_t \right] \leq \sqrt{2 \log(1/\delta)} \mathbb{E} \left[ \|A_*\|_{\bar{\Sigma}_t} \mid H_t \right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K \delta.$$

We also used that $A_t$ and $A_*$ are i.i.d. given $H_t$.

Since the above bound holds for any history $H_t$, we combine everything and get

$$\mathbb{E} \left[ \sum_{t=1}^{n} A_*^\top (\Theta_* - A_t^\top \Theta_t) \right] \leq \sqrt{2 \log(1/\delta)} \mathbb{E} \left[ \sum_{t=1}^{n} \|A_t\|_{\bar{\Sigma}_t} \right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K n \delta$$

$$\leq \sqrt{2 n \log(1/\delta)} \mathbb{E} \left[ \sum_{t=1}^{n} \|A_t\|_{\bar{\Sigma}_t}^2 \right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K n \delta$$

$$\leq \sqrt{2 n \log(1/\delta)} \mathbb{E} \left[ \sum_{t=1}^{n} \|A_t\|_{\bar{\Sigma}_t}^2 \right] + \sqrt{\frac{2}{\pi}} \sigma_{\max} K n \delta.$$

The second step uses the Cauchy-Schwarz inequality and the third step uses the concavity of the square root.

Since $a$ is an indicator vector, $\|A_t\|_{\bar{\Sigma}_t}^2 = \sigma_{t,A_t}^2$ is the marginal posterior variance of the mean reward of action $A_t$ in round $t$. Likewise, $\sigma_{\max}$ is the maximum marginal prior width of the mean reward at an action node. This concludes the proof.

B.2. Proof of Lemma 2

Since round $t$ is fixed, we write $L$ instead of $L_t$ and refer to node $\psi_t(i)$ by $i$ for any $i \in [L_t]$.

Fix any node $i$. By the total variance decomposition, where we introduce a parent parameter $\theta_{t,i-1}$,

$$\text{var} [\theta_{t,i} \mid H_t] = \mathbb{E} \left[ \sigma_{t,i}^2 \mid H_t \right] + \text{var} \left[ \hat{\theta}_{t,i} \mid H_t \right].$$
For Gaussian random variables, $\hat{\sigma}^2_{t,i}$ is independent of $\theta_{i,t-1}$, as shown in (7). Therefore,

$$\mathbb{E} \left[ \hat{\sigma}^2_{t,i} \mid H_t \right] = \hat{\sigma}^2_{t,i}. $$

For the second term, as shown in (7), $\hat{\theta}_{t,i} = \hat{\sigma}^2_{t,i} (\sigma_{0,i}^{-2} \theta_{i,t-1} + c)$, where $c$ is a constant conditioned on $H_t$. Therefore,

$$\text{var} \left[ \hat{\theta}_{t,i} \mid H_t \right] = \frac{\hat{\sigma}^4_{t,i}}{\sigma_{0,i}^4} \text{var} \left[ \theta_{i,t-1} \mid H_t \right].$$

Now we chain all identities for node $i$ and get

$$\text{var} \left[ \theta_{t,i} \mid H_t \right] = \hat{\sigma}^2_{t,i} + \frac{\hat{\sigma}^4_{t,i}}{\sigma_{0,i}^4} \text{var} \left[ \theta_{i,t-1} \mid H_t \right].$$

Finally, we apply the above identity recursively, from node $L$ up to the root, and get our claim,

$$\text{var} \left[ \theta_{t,L} \mid H_t \right] = \hat{\sigma}^2_{t,L} + \frac{\hat{\sigma}^4_{t,L}}{\sigma_{0,L}^4} \text{var} \left[ \theta_{t,L-1} \mid H_t \right] = \hat{\sigma}^2_{t,L} + \frac{\hat{\sigma}^4_{t,L}}{\sigma_{0,L}^4} \hat{\sigma}^2_{t,L-1} + \frac{\hat{\sigma}^4_{t,L}}{\sigma_{0,L}^4} \sigma_{0,L}^2 \text{var} \left[ \theta_{t,L-2} \mid H_t \right].$$

This completes the proof.

B.3. Proof of Lemma 3

Since round $t$ is fixed, we write $L$ instead of $L_i$ and refer to node $\psi_i(i)$ by $i$ for any $i \in \{L_i\}$.

For an action node $i$, the claim holds trivially since $\hat{\sigma}^2_{t+1,i} - \hat{\sigma}^2_{t,i} = \sigma^{-2}$. The rest of the proof is for internal nodes. We start with proving that

$$\sigma^{-2} \leq \hat{\sigma}^2_{t+1,i} - \hat{\sigma}^2_{t,i} \geq \sigma^{-2} \left( \prod_{j=i+1}^{L} \frac{\hat{\sigma}^4_{t,j}}{\sigma_{0,j}^4} \right) \frac{1}{1 + \sigma^{-2} \sigma_{0,i}^{-2}}$$

(15)

holds for any node $i$. The proof is by induction.

The basis of the induction is that the claim holds for action nodes. Let node $i = L$ be an action node, with $n$ observations by round $t$. We apply (8) and get

$$\hat{\sigma}^2_{t+1,i} - \hat{\sigma}^2_{t,i} = (\sigma_{0,i}^2 + (s + \varepsilon)(n + 1))^{-1} - (\sigma_{0,i}^2 + s/n)^{-1} = \sigma_{0,i}^{-2} [(\sigma_{0,i}^2 + s^{-2})^{-1} - (\sigma_{0,i}^2 + s^{-2}(n + 1))^{-1}]$$

$$= \sigma_{0,i}^{-4} (\sigma_{0,i}^2 + s^{-2})^{-2} \frac{\sigma^{-2}}{1 + \sigma^{-2} (\sigma_{0,i}^2 + s^{-2})^{-1}} = \frac{\hat{\sigma}^4_{t,i}}{\sigma_{0,i}^4} \frac{\sigma^{-2}}{1 + \sigma^{-2} \sigma_{0,i}^2}.$$}

The second and third equalities are by the Woodbury and Sherman-Morrison formulas, respectively, applied to scalars. The lower bound follows from $\hat{\sigma}^2_{t,i} / \sigma_{0,i}^2 \leq 1$ and $1 + \sigma^{-2} \sigma_{0,i}^{-2} \geq 1$.

In the inductive step, we assume that (15) holds for node $i + 1$ and prove it for node $i$. Note that node $i + 1$ is the only child of node $i$ where the posterior between rounds $t$ and $t + 1$ changes. Let $s = \sum_{j \in \psi_i(1)} \sigma_{0,j}^2$ and $\varepsilon = \sigma_{0,i}^{-2} - \sigma_{0,i}^{-2}$. Now we apply (9) and get

$$\hat{\sigma}_{t+1,i} - \hat{\sigma}_{t,i}^2 = (\sigma_{0,i}^2 + (s + \varepsilon)(n + 1))^{-1} - (\sigma_{0,i}^2 + s^{-1})^{-1} = \sigma_{0,i}^{-2} [(\sigma_{0,i}^2 + s)^{-1} - (\sigma_{0,i}^2 + s + \varepsilon)^{-1}]$$

$$= \sigma_{0,i}^{-4} (\sigma_{0,i}^2 + s)^{-2} \frac{\varepsilon}{1 + (\sigma_{0,i}^2 + s)^{-1} \varepsilon} = \frac{\hat{\sigma}^4_{t,i}}{\sigma_{0,i}^4} \frac{\varepsilon}{1 + \sigma_{0,i}^2 \varepsilon}.$$}

As in the proof for the action node, the second and third equalities are by the Woodbury and Sherman-Morrison formulas, respectively, applied to scalars. The lower bound follows from $\hat{\sigma}^2_{t,i} / \sigma_{0,i}^2 \leq 1$, $1 + \hat{\sigma}^2_{t,i} \varepsilon \geq 1$, and $\varepsilon \leq \sigma^{-2}$. 

To complete the proof, we have from (7) that
\[ \hat{\sigma}_{t+1,i}^{-2} - \hat{\sigma}_{t,i}^{-2} = \bar{\sigma}_{t+1,i}^{-2} - \bar{\sigma}_{t,i}^{-2} \]
holds for any internal node \( i \). Finally, note that \( 1 + \sigma^{-2} \hat{\sigma}_{0,i}^{-2} \leq 1 + \sigma^{-2} \sigma_{0,\max}^{-2} = c. \)

**B.4. Proof of Theorem 4**

First, we apply Lemma 1 and get that
\[ BR(n) \leq \sqrt{2nV(n) \log(1/\delta)} + \sqrt{2/\pi \sigma_{\max} K n \delta}, \]
where \( V(n) = E \left[ \sum_{t=1}^{n} \bar{\sigma}_{t,A_t}^{2} \right] \) and \( \bar{\sigma}_{t,A_t}^{2} \) is the marginal posterior variance in node \( A_t \) in round \( t \). We derive a worst-case upper bound on \( \mathcal{V}(n) \) next.

We start with a worst-case upper in any round \( t \). Since round \( t \) is fixed, we write \( L \) instead of \( L_t \) and refer to node \( \psi_{i}(t) \) by \( i \) for any \( i \in [L_t] \). Let \( s_i = \prod_{j=\lambda+1}^{L} \frac{\hat{\sigma}_{t,i}^{2}}{\sigma_{\tilde{\sigma},j}^{2}} \). Then by Lemma 2,
\[ \frac{\hat{\sigma}_{t,A_t}^{2}}{\sigma^{2}} = \hat{\sigma}_{t,i}^{2} \leq \frac{\hat{\sigma}_{t,i}^{2}}{\sigma^{2}} \leq \frac{c \log(1 + s_i \hat{\sigma}_{t,i}^{2})}{\sigma^{2}} \leq \frac{c \log(1 + \sigma_{0,\hat{\sigma}}^{-2})}{\sigma^{2}} = c_i. \]
The first inequality holds because \( s_i \leq 1 \), and \( ax / \log(1 + ax) \leq x / \log(1 + x) \) for any \( a \in [0, 1] \) and \( x > 0 \). The second inequality holds because \( x / \log(1 + x) \) is maximized when \( x \) is, which happens at \( \hat{\sigma}_{t,i} = \sigma_{0,i} \).

For any \( c \geq 1 \) and node \( i \in [L], \)
\[ \log \left( 1 + s_i \hat{\sigma}_{t,i}^{2} \right) = c^{L-i} \log \left( 1 + s_i \hat{\sigma}_{t,i}^{2} \right) \leq c^{L-i} \log \left( 1 + \sigma_{0}^{2} \hat{\sigma}_{t,i}^{2} \right), \]
where the inequality holds because \( a \log(1 + x) \leq \log(1 + ax) \) for any \( a \in [0, 1] \) and \( x > 0 \). Moreover,
\[ \log \left( 1 + c^{L-i} s_i \hat{\sigma}_{t,i}^{2} \right) = \log \left( \hat{\sigma}_{t,i}^{-2} + c^{L-i} s_i \hat{\sigma}_{t,i}^{2} \right) \leq \log(\hat{\sigma}_{t,i}^{-2}) - \log(\hat{\sigma}_{t+1,i}^{-2}), \]
where the last step is by Lemma 3. Now we chain all inequalities, switch to the full notation, and get
\[ \hat{\sigma}_{t,A_t}^{2} \leq \sigma_{0}^{2} \sum_{i=1}^{L} c^{L-i} \left[ \log(\hat{\sigma}_{t+1,i}^{2} - \log(\sigma_{0,i}^{2}) \right] \leq \sigma_{0}^{2} \sum_{i=1}^{L} c^{L-i} \log(\hat{\sigma}_{t+1,i}^{2} - \log(\sigma_{0,i}^{2})). \]
Finally, we sum up the above upper bound over all rounds \( t \). Let \( h_i \) be the maximum length of any path from node \( i \) to its descendant. Due to \( c \geq 1 \) and telescoping in the above decomposition, we get
\[ \mathcal{V}(n) \leq \sigma_{0}^{2} \sum_{i \in \mathcal{N}} c^{h_i} \left[ \log(\hat{\sigma}_{n+1,i}^{2} - \log(\sigma_{0,i}^{2}) \right] = \sigma_{0}^{2} \sum_{i \in \mathcal{N}} c^{h_i} \log(\hat{\sigma}_{0,i}^{2} - \sigma_{0,0}^{2} \sigma_{n+1,i}^{2}). \]
For an action node \( i, \hat{\sigma}_{n+1,i}^{2} \leq \sigma_{0,i}^{2} + \sigma^{-2} n, \) and thus
\[ \log(\hat{\sigma}_{0,i}^{2} - \sigma_{0,0}^{2}) \leq \log \left( 1 + \frac{\sigma_{0,i}^{2} n}{\sigma^{2}} \right). \]
For an internal node $i$, $\sigma_{n+1,i}^{-2} \leq \sigma_{0,i}^{-2} + \sum_{j \in \text{ch}(i)} \sigma_{0,j}^{-2}$, and thus

$$\log(\sigma_{0,j}^{-2}, \sigma_{n+1,i}^{-2}) \leq \log \left( 1 + \sigma_{0,j}^{-2} \sum_{j \in \text{ch}(i)} \sigma_{0,j}^{-2} \right).$$

This completes the proof.

C. Contextual Regret Bound Proofs

This section contains proofs of our contextual regret bound and supporting lemmas.

C.1. Proof of Theorem 5

First, we apply Lemma 1 and get that

$$BR(n) \leq \sqrt{2n V(n) \log(1/\delta)} + \sqrt{2/\pi \sigma_{\max} K n \delta},$$

where $V(n) = \mathbb{E} \left[ \sum_{i=1}^{n} \tilde{\sigma}_{t,A_i}^2 \right]$ and $\tilde{\sigma}_{t,A_i}^2 = ||X_t||_{\Sigma_t,A_i}$ denotes the marginal posterior variance in node $A_t$ in round $t$, in the direction of context $X_t$. We derive a worst-case upper bound on $V(n)$ next.

We start with a worst-case upper in any round $t$. Since round $t$ is fixed, we write $L$ instead of $L_t$ and refer to node $\psi_t(i)$ by $i$ for any $i \in [L_t]$. Let $S_i = \prod_{j=i+1}^{L} \Sigma_{0,j}^{-1} \Sigma_{t,j}$. Then by Lemma 8,

$$\sigma_{t,A_t}^2 = X_t^T \tilde{\Sigma}_{t,A_t} X_t = \sigma^2 \sum_{i=1}^{L} \sigma^{-2} X_i^T \tilde{\Sigma}_{t,i} S_i X_t \leq \sigma^2 \sum_{i=1}^{L} c_i \log \left( 1 + \sigma^{-2} X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right),$$

where $c_i$ is an upper bound at node $i$ defined as

$$\frac{X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t}{\sigma^2 \log \left( 1 + \sigma^{-2} X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right)} \leq \frac{\lambda_1 \left( \Sigma_{t,i} \right)}{\sigma^2 \log \left( 1 + \sigma^{-2} \lambda_d \left( \Sigma_{t,i} \right) \right)} \leq \frac{\sigma_{0,i}^2}{\sigma^2 \log \left( 1 + \frac{\sigma_{0,i}^2}{\sigma^2} \right)} = c_i.$$

The first inequality holds because $\lambda_1(S_t) \leq 1$ and $||X_t||_2 \leq 1$, where the former follows from the definitions of $\tilde{\Sigma}_{t,j}$ and $\Sigma_{0,j} = \sigma_{0,j}^2 I_d$. We also use the fact that $ax / \log(1 + ax) \leq x / \log(1 + x)$ holds for any $a \in [0, 1]$ and $x > 0$. The second inequality holds because $\lambda_1 \left( \Sigma_{t,i} \right) \leq \lambda_1 \left( \Sigma_{0,i} \right) = \sigma_{0,i}^2$. We also use that $x / \log(1 + x)$ is maximized when $x = 1$.

For any $c \geq 1$ and node $i \in [L]$,

$$\log \left( 1 + \sigma^{-2} X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right) = c^{L_i} \log \left( 1 + \sigma^{-2} X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right) \leq c^{L_i} \log \left( 1 + \sigma^{-2} X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right),$$

where the inequality holds because $a \log(1 + x) \leq \log(1 + ax)$ for any $a \in [0, 1]$ and $x > 0$. Moreover,

$$\log \left( 1 + \sigma^{-2} c^{L_i} X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right) = \log \left( 1 + \sigma^{-2} c^{L_i} \tilde{\Sigma}_{t,i} S_i X_t X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right) = \log \left( \tilde{\Sigma}_{t,i}^{-1} + \sigma^{-2} c^{L_i} S_i X_t X_i^T S_i^T \tilde{\Sigma}_{t,i} S_i X_t \right) - \log \left( \tilde{\Sigma}_{t,i}^{-1} \right) \leq \log \left( \tilde{\Sigma}_{t,i}^{-1} \right),$$

where the last step is by Lemma 9. Now we chain all inequalities, switch to the full notation, and get

$$\sigma_{t,A_t}^2 \leq \sigma^2 \sum_{i=1}^{L_i} c^{L_i} \log \left( \tilde{\Sigma}_{t+1}^{-1} \psi_t(i) \right) - \log \left( \tilde{\Sigma}_{t+1}^{-1} \psi_t(i) \right).$$

Finally, we sum up the above upper bound over all rounds $t$. Let $h_i$ be the maximum length of any path from node $i$ to its descendant. Due to $c \geq 1$ and telescoping in the above decomposition, we get

$$V(n) \leq \sigma^2 \sum_{i \in \mathcal{N}} c^{h_i} c_i \left[ \log \left( \tilde{\Sigma}_{n+1,i}^{-1} \psi_t(i) \right) - \log \left( \tilde{\Sigma}_{0,i}^{-1} \psi_t(i) \right) \right] = \sigma^2 \sum_{i \in \mathcal{N}} c^{h_i} c_i \log \left( \tilde{\Sigma}_{0,i}^{-1} \tilde{\Sigma}_{n+1,i}^{-1} \psi_t(i) \right).$$
We start with proving that for any node $i$,

$$\log \det (\Sigma_{0,i}^{-1} S_{n+1,i}^{-1} \Sigma_{0,i}^{-1}) \leq d \log \left( \frac{1}{d} \text{tr}(\Sigma_{0,i}^{-1} S_{n+1,i}^{-1} \Sigma_{0,i}^{-1}) \right) \leq d \log \left( 1 + \frac{\sigma_{0,i}^2}{d} \right).$$

The lower bound follows from $\Sigma$. The second and third equalities are by the Woodbury and Sherman-Morrison formulas, respectively. The matrix is rank $1$, so

$$\log \det (\Sigma_{0,i}^{-1} S_{n+1,i}^{-1} \Sigma_{0,i}^{-1}) \leq d \log \left( 1 + \sigma_{0,i}^2 \sum_{j \in \text{eh}(i)} \sigma_{o,j}^2 \right).$$

This completes the proof.

**C.2. Supporting Lemmas**

The first claim is a matrix generalization of Lemma 2.

**Lemma 8.** In any round $t$, the marginal posterior covariance in action node $A_t$ decomposes as

$$\hat{\Sigma}_{t,A_t} = \sum_{i=1}^{L_t} S_i^T \hat{\Sigma}_{t,\psi(i)} S_i,$$

where $S_i = \prod_{j=i+1}^{L_t} \Sigma_{0,\psi(j)}^{-1} \hat{\Sigma}_{t,\psi(j)}$. The rest of the proof is for internal nodes.

**Proof.** The proof is exactly the same as in Appendix B.2, except that the variances are replaced by covariances.

The second claim is a matrix generalization of Lemma 3.

**Lemma 9.** Fix any round $t$ and $i \in [L_t]$. Then

$$\hat{\Sigma}_{t+1,\psi(i)} - \hat{\Sigma}_{t,\psi(i)} \succeq \sigma^{-2} c^{i-L_t} S_i X_i X_i^T S_i^{-1},$$

where $c = 1 + \sigma_{0,\text{max}}^2 / \sigma^2$ and $S_i$ is defined in Lemma 8.

**Proof.** Since round $t$ is fixed, we write $L$ instead of $L_t$ and refer to node $\psi(i)$ by $i$ for any $i \in [L_t]$.

For an action node $i$, the claim holds trivially since $\hat{\Sigma}_{t+1,i}^{-1} - \hat{\Sigma}_{t,i}^{-1} = \sigma^{-2} X_i X_i^T$. The rest of the proof is for internal nodes. For any node $i$, $\hat{\Sigma}_{t+1,i}^{-1} - \hat{\Sigma}_{t,i}^{-1} = \nu \nu^T$ is a rank-1 matrix for some $\nu \in \mathbb{R}^d$, such that

$$\nu \nu^T \succeq \sigma^{-2} c^{i-L_t} S_i X_i X_i^T S_i^{-1}, \quad \nu^T \nu \leq \sigma^{-2}. \quad (16)$$

The proof is by induction.

The basis of the induction is that the claim holds for action nodes. Let node $i = L$ be an action node. We apply (11) and get

$$\hat{\Sigma}_{t+1,i}^{-1} - \hat{\Sigma}_{t,i}^{-1} = (\Sigma_{0,i} + (G_{t,i} + \sigma^{-2} X_i X_i^T)^{-1})^{-1} - (\Sigma_{0,i} + G_{t,i})^{-1}$$

$$= \Sigma_{0,i}^{-1} [(\Sigma_{0,i} + G_{t,i})^{-1} - (\Sigma_{0,i} + G_{t,i} + \sigma^{-2} X_i X_i^T)^{-1} \Sigma_{0,i}]$$

$$= \Sigma_{0,i}^{-1} (\Sigma_{0,i} + G_{t,i})^{-1} - (\Sigma_{0,i} + G_{t,i} + \sigma^{-2} X_i X_i^T)^{-1} \Sigma_{0,i}$$

$$= \Sigma_{0,i}^{-1} \frac{\sigma^{-2} X_i X_i^T}{1 + \sigma^{-2} X_i^T \Sigma_{0,i}^{-1} X_i} \hat{\Sigma}_{t,i} \Sigma_{0,i}^{-1}.$$

The second and third equalities are by the Woodbury and Sherman-Morrison formulas, respectively. The matrix is rank $1$. The lower bound follows from $\Sigma_{0,i} = \sigma_{0,i}^2 I_d$ and $\|X_i\|_2 \leq 1$, which can be used to derive

$$1 + \sigma^{-2} X_i^T \hat{\Sigma}_{t,i} X_i \leq 1 + \sigma^{-2} \lambda_1 \left( \hat{\Sigma}_{t,i} \right) \leq 1 + \sigma^{-2} \sigma_{0,i}^2 \leq c.$$
The upper bound uses that $\|\Sigma_{0,i}^{-1}\hat{\Sigma}_{t,i}X_t\|_2 \leq 1$ and $1 + x \geq 1$ for any $x \geq 0$.

In the inductive step, we assume that (16) holds for node $i + 1$ and prove it for node $i$. Note that node $i + 1$ is the only child of node $i$ where the posterior between rounds $t$ and $t + 1$ changes. Let $M = \sum_{j \in \text{ch}(i)} \hat{\Sigma}_{t,j}$ and $\varepsilon = \Sigma_{t+1,i+1} - \hat{\Sigma}_{t,i+1}$. Now we apply (12) and get

$$
\hat{\Sigma}_{t+1,i} - \hat{\Sigma}_{t,i}^{-1} = (\Sigma_{0,i} + (M + \varepsilon)^{-1})^{-1} - (\Sigma_{0,i}^{-1} + M^{-1})^{-1}
$$

$$
= \Sigma_{0,i}^{-1}(\Sigma_{0,i}^{-1} + M)^{-1} - (\Sigma_{0,i}^{-1} + M + \varepsilon)^{-1} \Sigma_{0,i}^{-1}
$$

$$
= \Sigma_{0,i}^{-1}(\Sigma_{0,i}^{-1} + M)^{-1}\frac{\varepsilon}{1 + v^T(\Sigma_{0,i}^{-1} + M)^{-1}v} (\Sigma_{0,i}^{-1} + M)^{-1} \Sigma_{0,i}^{-1}
$$

$$
= \Sigma_{0,i}^{-1}\hat{\Sigma}_{t,i}^{-1} v v^T (1 + v^T\hat{\Sigma}_{t,i}^{-1}v)^{-1} \hat{\Sigma}_{t,i}^{-1} \Sigma_{0,i}^{-1}
$$

where $\varepsilon = vv^T$. The matrix is rank 1, because by the inductive argument $\varepsilon$ is rank 1. As in the proof for the action node, the second and third equalities are by the Woodbury and Sherman-Morrison formulas. The lower bound is derived using

$$
1 + v^T \hat{\Sigma}_{t,i} v \leq 1 + \|v\|^2 \lambda_1(\hat{\Sigma}_{t,i}) \leq 1 + \sigma_{0,i}^2 \leq c.
$$

This follows from $\Sigma_{0,i} = \sigma_{0,i}^2 I_d$ and $v^T v \leq \sigma^{-2}$, where the latter is by the inductive argument. The upper bound uses that $\|\Sigma_{0,i}^{-1}\hat{\Sigma}_{t,i}v\|_2 \leq \sigma^{-1}$ and $1 + x \geq 1$ for any $x \geq 0$.

To complete the proof, we have from (10) that

$$
\hat{\Sigma}_{t+1,i} - \hat{\Sigma}_{t,i}^{-1} = \hat{\Sigma}_{t+1,i+1} - \hat{\Sigma}_{t,i+1}
$$

holds for any internal node $i$. □