Hierarchical Bayesian Bandits

Joey Hong
UC Berkeley*

Branislav Kveton
Amazon*

Manzil Zaheer
Google DeepMind

Mohammad Ghavamzadeh
Google Research

Abstract

Meta-, multi-task, and federated learning can be all viewed as solving similar tasks, drawn from a distribution that reflects task similarities. We provide a unified view of all these problems, as learning to act in a *hierarchical Bayesian bandit*. We propose and analyze a natural hierarchical Thompson sampling algorithm (`HierTS`) for this class of problems. Our regret bounds hold for many variants of the problems, including when the tasks are solved sequentially or in parallel; and show that the regret decreases with a more informative prior. Our proofs rely on a novel total variance decomposition that can be applied beyond our models. Our theory is complemented by experiments, which show that the hierarchy helps with knowledge sharing among the tasks. This confirms that hierarchical Bayesian bandits are a universal and statistically-efficient tool for learning to act with similar bandit tasks.

1 INTRODUCTION

A *stochastic bandit* (Lai and Robbins, 1985; Auer et al., 2002; Lattimore and Szepesvari, 2019) is an online learning problem where a *learning agent* sequentially interacts with an environment over n rounds. In each round, the agent takes an *action* and receives a *stochastic reward*. 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 must learn them by taking the actions. This induces the *exploration-exploitation dilema*: *explore*, and learn more about an action; or *exploit*, and take the action with the highest estimated reward. In online advertising, an action could be showing an advertisement and its reward could be an indicator of a click.

More statistically-efficient exploration is the primary topic of bandit papers. This is attained by leveraging the structure of the problem, such as the form of the reward distribution (Garivier and Cappe, 2011), prior distribution over model parameters (Thompson, 1933; Agrawal and Goyal, 2012; Chapelle and Li, 2012; Russo et al., 2018), conditioning on known feature vectors (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013), or modeling the process by which the total reward arises (Radlinski et al., 2008; Kveton et al., 2015a; Gai et al., 2012; Chen et al., 2016; Kveton et al., 2015b). In this work, we solve multiple similar bandit tasks, and each task teaches the agent how to solve other tasks more efficiently.

We formulate the problem of learning to solve similar bandit tasks as regret minimization in a *hierarchical Bayesian model* (Gelman et al., 2013). Each task is parameterized by a *task parameter*, which is sampled i.i.d. from a distribution parameterized by a *hyper-parameter*. The parameters are unknown and this relates all tasks, in the sense that each task teaches the agent about any other task. We derive Bayes regret bounds that reflect the structure of the problem and show that the price for learning the hyper-parameter is low. Our derivations use a novel *total variance decomposition*, which decomposes the parameter uncertainty into per-task uncertainty conditioned on knowing the hyper-parameter and hyper-parameter uncertainty. After that, we individually bound each uncertainty source by elliptical lemmas (Dani et al., 2008; Abbasi-Yadkori et al., 2011). Our approach can be exactly implemented and analyzed in hierarchical multi-armed and linear bandits with Gaussian rewards, but can be extended to other graphical model structures.

We build on numerous prior works that study a similar structure, under the names of collaborating filtering bandits (Gentile et al., 2014; Kawale et al., 2015; Li et al., 2016), bandit meta-learning and multi-task learning (Azar et al., 2013; Deshmukh et al., 2017; Bas-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

*The work started while being at Google Research.

tani et al., 2019; Cella et al., 2020; Kveton et al., 2021; Moradipari et al., 2021), and representation learning (Yang et al., 2021). Despite it, we make major novel contributions, both in terms of a more general setting and analysis techniques. Our setting relaxes the assumptions that the tasks are solved in a sequence and that exactly one task is solved per round. Moreover, while the design of our posterior sampling algorithm is standard, we make novel contributions in its analysis. In the sequential setting (Section 6.1), we derive a Bayes regret bound by decomposing the posterior covariance, which is an alternative to prior derivations based on filtered mutual information (Russo and Van Roy, 2016; Lu and Van Roy, 2019). This technique is general, simple, and yields tighter regret bounds because it avoids marginal task parameter covariance; and so is of a broad interest. In the concurrent setting (Section 6.3), we bound the additional regret due to not updating the posterior after each interaction. This is non-trivial and a major departure from other bandit analyses. Our Bayes regret bound for this setting is the first of its kind.

The paper is organized as follows. In Section 2, we formalize our setting of *hierarchical Bayesian bandits*. In Section 3, we introduce a natural Thompson sampling algorithm (`HierTS`) for solving it. In Section 4, we instantiate it in hierarchical Gaussian models. In Section 5, we review key ideas in our regret analyses, including a novel total covariance decomposition that allows us to analyze posteriors in hierarchical models. In Section 6, we prove Bayes regret bounds for `HierTS` in sequential and concurrent settings. Finally, in Section 7, we evaluate `HierTS` empirically to confirm our theoretical results.

2 SETTING

We use the following notation. Random variables are capitalized, except for Greek letters like θ and μ. For any positive integer n, we define $[n] = \{1, \ldots, n\}$. The indicator function is denoted by $\mathbb{1}\{\cdot\}$. The i-th entry of vector v is v_i. If the vector is already indexed, such as v_j, we write $v_{j,i}$. A matrix with diagonal entries v is $\mathrm{diag}\,(v)$. For any matrix $M \in \mathbb{R}^{d \times d}$, the maximum eigenvalue is $\lambda_1(M)$ and the minimum is $\lambda_d(M)$. The big O notation up to logarithmic factors is $\tilde{O}$.

Now we present our setting for solving similar bandit tasks. Each task is a *bandit instance* with actions $a \in \mathcal{A}$, where $\mathcal{A}$ denotes an *action set*. Rewards of actions are generated by *reward distribution* $P(\cdot \mid a; \theta)$, where $\theta \in \Theta$ is an unknown parameter shared by all actions. We assume that the rewards are σ^2-sub-Gaussian and denote by $r(a; \theta) = \mathbb{E}_{Y \sim P(\cdot \mid a; \theta)}[Y]$ the mean reward of action a under θ. The learning agent interacts with m tasks. In a recommender system, each task could be an individual user. The task $s \in [m]$ is parameterized by a *task parameter* $\theta_{s,*} \in \Theta$, which is sampled i.i.d. from a *task prior distribution* $\theta_{s,*} \sim P(\cdot \mid \mu_*)$; which is parameterized by an unknown *hyper-parameter* μ_*.

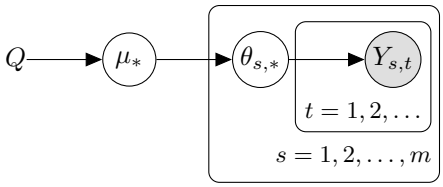

Figure 1: Graphical model of our hierarchical Bayesian bandit.

The agent acts at discrete decision points, which are integers and we call them *rounds*. At round $t \geq 1$, the agent is asked to act in a set of tasks $\mathcal{S}_t \subseteq [m]$. It takes actions $A_t = (A_{s,t})_{s \in \mathcal{S}_t}$, where $A_{s,t} \in \mathcal{A}$ is the action in task s; and receives rewards $Y_t = (Y_{s,t})_{s \in \mathcal{S}_t} \in \mathbb{R}^{|\mathcal{S}_t|}$, where $Y_{s,t} \sim P(\cdot \mid A_{s,t}; \theta_{s,*})$ is a stochastic reward for taking action $A_{s,t}$ in task s. The rewards are drawn i.i.d. from their respective distributions. The set $\mathcal{S}_t$ can depend arbitrarily on the history. The assumption that the action set $\mathcal{A}$ is the same across all tasks and rounds is only to simplify exposition.

In *hierarchical Bayesian bandits*, the hyper-parameter μ_* is initially sampled from a *hyper-prior* Q known by the learning agent. Our full model is given by

$$\begin{aligned} \mu_* &\sim Q\,, \\ \theta_{s,*} \mid \mu_* &\sim P(\cdot \mid \mu_*)\,, && \forall s \in [m]\,, \\ Y_{s,t} \mid A_{s,t}, \theta_{s,*} &\sim P(\cdot \mid A_{s,t}; \theta_{s,*})\,, && \forall t \geq 1,\ s \in \mathcal{S}_t\,, \end{aligned}$$

and also visualized in Figure 1. Note that P denotes both the task prior and reward distribution; but they can be distinguished based on their parameters. Our setting is an instance of a hierarchical Bayesian model commonly used in supervised learning (Lindley and Smith, 1972; Zhang and Yang, 2017), and has been studied in bandits in special cases where tasks appear sequentially (Kveton et al., 2021; Basu et al., 2021).

Our learning agent interacts with each of m tasks for at most n times. So the total number of rounds varies, as it depends on the number of tasks that the agent interacts with simultaneously in each round. However, the maximum number of interactions is mn. The goal is to minimize the *Bayes regret* (Russo and Van Roy, 2014) defined as

$$\mathcal{BR}(m, n) = \mathbb{E}\left[\sum_{t \geq 1} \sum_{s \in \mathcal{S}_t} r(A_{s,*}; \theta_{s,*}) - r(A_{s,t}; \theta_{s,*})\right]\,,$$

where $A_{s,*} = \arg\max_{a \in \mathcal{A}} r(a; \theta_{s,*})$ is the optimal action in task s. The expectation in $\mathcal{BR}(m, n)$ is over μ_*,

Algorithm 1 Hierarchical Thompson sampling.

Input: Hyper-prior Q
Initialize $Q_1 \leftarrow Q$
for $t = 1, 2, \ldots$ **do**
  Observe tasks $\mathcal{S}_t \subseteq [m]$
  Sample $\mu_t \sim Q_t$
  for $s \in \mathcal{S}_t$ **do**
    Compute $P_{s,t}(\theta \mid \mu_t) \propto \mathcal{L}_{s,t}(\theta) P(\theta \mid \mu_t)$
    Sample $\theta_{s,t} \sim P_{s,t}(\cdot \mid \mu_t)$
    Take action $A_{s,t} \leftarrow \arg\max_{a \in \mathcal{A}} r(a; \theta_{s,t})$
    Observe reward $Y_{s,t}$
  Update Q_{t+1}

$\theta_{s,*}$ for $s \in [m]$, and actions $A_{s,t}$ of the agent. While weaker than a traditional frequentist regret, the Bayes regret is a practical performance metric, as we are often interested in an average performance (Hong et al., 2020; Kveton et al., 2021). For example, when recommending to a group of users, it is natural to optimize over the whole population rather than an individual. Our definition of $\mathcal{BR}(m, n)$ is also dictated by the fact that m and n are the primary quantities of interest in our regret analyses. Our goal is to minimize $\mathcal{BR}(m, n)$ without knowing μ_* and $\theta_{s,*}$ a priori.

Since the set of tasks $\mathcal{S}_t$ can be chosen arbitrarily, our setting is general and subsumes many prior settings. For instance, when the agent interacts with the same task for n rounds before shifting to the next one, and $\mathcal{S}_t = \{\lceil t/n \rceil\}$, we get a *meta-learning bandit* (Kveton et al., 2021). More generally, when the agent interacts with the tasks sequentially, $|\mathcal{S}_t| = 1$, our setting can be viewed as multi-task learning where any task helps the agent to solve other tasks. Therefore, we recover a *multi-task bandit* (Wan et al., 2021). Finally, when the agent acts in multiple tasks concurrently, $|\mathcal{S}_t| > 1$, we recover the setting of *collaborative filtering bandits* (Gentile et al., 2014; Li et al., 2016) or more recently *federated bandits* (Shi and Shen, 2021). Our algorithm and its analysis apply to all of these settings.

3 ALGORITHM

We take a Bayesian view and use hierarchical *Thompson sampling (TS)*, which we call `HierTS`, to solve our problem class. `HierTS` samples task parameters from their posterior conditioned on history. Specifically, let $H_{s,t} = ((A_{s,\ell}, Y_{s,\ell}))_{\ell < t,\, s \in \mathcal{S}_\ell}$ denote the history of all interactions of `HierTS` with task s until round t, and $H_t = (H_{s,t})_{s \in [m]}$ be the concatenation of all histories up to round t. For each task $s \in \mathcal{S}_t$ in round t, `HierTS` samples $\theta_{s,t} \sim \mathbb{P}(\theta_{s,*} = \cdot \mid H_t)$ and then takes action $A_{s,t} = \arg\max_{a \in \mathcal{A}} r(a; \theta_{s,t})$. The key difference from classical Thompson sampling is that the history H_t includes observations of multiple tasks.

To sample $\theta_{s,t}$, we must address how the uncertainty over the unknown hyper-parameter μ_* and task parameters $\theta_{s,*}$ is modeled. The key idea is to maintain a *hyper-posterior* Q_t over μ_*, given by

$$Q_t(\mu) = \mathbb{P}(\mu_* = \mu \mid H_t) \,,$$

and then perform two-stage sampling. In particular, in round t, we first sample hyper-parameter $\mu_t \sim Q_t$. Next, for any task $s \in \mathcal{S}_t$, we sample the task parameter $\theta_{s,t} \sim P_{s,t}(\cdot \mid \mu_t)$, where

$$P_{s,t}(\theta \mid \mu) = \mathbb{P}(\theta_{s,*} = \theta \mid \mu_* = \mu, H_{s,t}) \,.$$

In $P_{s,t}(\cdot \mid \mu)$, we only condition on the history of task s, since $\theta_{s,*}$ is independent of the other task histories given $\mu_* = \mu$ (Figure 1). This process clearly samples from the true posterior, which is given by

$$\begin{aligned} \mathbb{P}(\theta_{s,*} = \theta \,|\, H_t) &= \int_\mu \mathbb{P}(\theta_{s,*} = \theta, \mu_* = \mu \,|\, H_t) \,\mathrm{d}\mu \\ &= \int_\mu P_{s,t}(\theta \mid \mu) Q_t(\mu) \,\mathrm{d}\mu \,, \end{aligned} \tag{1}$$

where $P_{s,t}(\theta \mid \mu) \propto \mathcal{L}_{s,t}(\theta) P(\theta \mid \mu)$ and

$$\mathcal{L}_{s,t}(\theta) = \textstyle\prod_{(a,y) \in H_{s,t}} P(y \mid a; \theta)$$

denotes the likelihood of rewards in task s given task parameter θ.

The pseudo-code of `HierTS` is shown in Algorithm 1. Sampling in (1) can be implemented exactly in Gaussian graphical models (Section 4). These models have interpretable closed-form posteriors, which permit the regret analysis of `HierTS`. In practice, `HierTS` can be implemented for any posterior distributions, but may require approximate inference (Doucet et al., 2001) to tractably sample from the posterior.

4 HIERARCHICAL GAUSSIAN BANDITS

Now we instantiate `HierTS` in hierarchical Gaussian models. This yields closed-form posteriors, which permit regret analysis (Section 5). We discuss generalization to other distributions in Section 5.4.

We assume that the environment is generated as

$$\begin{aligned} \mu_* &\sim \mathcal{N}(\mu_q, \Sigma_q) \,, && \\ \theta_{s,*} \mid \mu_* &\sim \mathcal{N}(\mu_*, \Sigma_0) \,, && \forall s \in [m] \,, \\ Y_{s,t} \mid A_{s,t}, \theta_{s,*} &\sim \mathcal{N}(A_{s,t}^\top \theta_{s,*}, \sigma^2) \,, && \forall t \geq 1,\, s \in \mathcal{S}_t \,, \end{aligned} \tag{2}$$

where $\Sigma_q \in \mathbb{R}^{d \times d}$ and $\Sigma_0 \in \mathbb{R}^{d \times d}$ are covariance matrices; μ_q, μ_*, $\theta_{s,*}$ are d-dimensional vectors; the set

of actions is $\mathcal{A} \subseteq \mathbb{R}^d$; and the mean reward of action $a \in \mathcal{A}$ is $r(a;\theta) = a^\top \theta$. The reward noise is $\mathcal{N}(0, \sigma^2)$. This formulation captures both the multi-armed and linear bandits, since the actions in the former can be viewed as vectors in a standard Euclidean basis. We assume that all of μ_q, Σ_q, Σ_0, and σ are known by the agent. This assumption is only needed in the analysis of `HierTS`, where we require an analytically tractable posterior. We relax it in our experiments (Section 7), where we learn these quantities from past data.

4.1 Gaussian Bandit

We start with a K-armed Gaussian bandit, which we instantiate as (2) as follows. The task parameter $\theta_{s,*}$ is a vector of mean rewards in task s, where $\theta_{s,*,i}$ is the mean reward of action i. The covariance matrices are diagonal, $\Sigma_q = \sigma_q^2 I_K$ and $\Sigma_0 = \sigma_0^2 I_K$. We assume that both $\sigma_q > 0$ and $\sigma_0 > 0$ are known. The reward distribution of action i is $\mathcal{N}(\theta_{s,*,i}, \sigma^2)$, where $\sigma > 0$ is a known reward noise.

Because Σ_q and Σ_0 are diagonal, the hyper-posterior in round t factors across the actions. Specifically, it is $Q_t = \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t)$, where $\bar{\Sigma}_t = \mathrm{diag}\left((\bar{\sigma}_{t,i}^2)_{i \in [K]}\right)$ and

$$\bar{\mu}_{t,i} = \bar{\sigma}_{t,i}^2 \left(\frac{\mu_{q,i}}{\sigma_q^2} + \sum_{s \in [m]} \frac{N_{s,t,i}}{N_{s,t,i}\sigma_0^2 + \sigma^2} \frac{B_{s,t,i}}{N_{s,t,i}} \right) , \quad (3)$$
$$\bar{\sigma}_{t,i}^{-2} = \sigma_q^{-2} + \sum_{s \in [m]} \frac{N_{s,t,i}}{N_{s,t,i}\sigma_0^2 + \sigma^2} .$$

Here $N_{s,t,i} = \sum_{\ell < t} \mathbb{1}\{s \in \mathcal{S}_\ell, A_{s,\ell} = i\}$ is the number of times that action i is taken in task s up to round t and $B_{s,t,i} = \sum_{\ell < t} \mathbb{1}\{s \in \mathcal{S}_\ell, A_{s,\ell} = i\} Y_{s,\ell}$ is its total reward. The hyper-posterior is derived in Appendix D of Kveton et al. (2021). To understand it, it is helpful to view it as a Gaussian posterior where each task is a single observation. The observation of task s is the empirical mean reward estimate of action i in task s, $B_{s,t,i}/N_{s,t,i}$, and its variance is $(N_{s,t,i}\sigma_0^2 + \sigma^2)/N_{s,t,i}$. The tasks with more observations affect the value of $\bar{\mu}_{t,i}$ more, because their mean reward estimates have lower variances. The variance never decreases below σ_0^2, because even the actual mean reward $\theta_{s,*,i}$ would be a noisy observation of $\mu_{*,i}$ with variance σ_0^2.

After the hyper-parameter is sampled, $\mu_t \sim Q_t$, the task parameter is sampled, $\theta_{s,t} \sim \mathcal{N}(\tilde{\mu}_{s,t}, \tilde{\Sigma}_{s,t})$, where $\tilde{\Sigma}_{s,t} = \mathrm{diag}\left((\tilde{\sigma}_{s,t,i}^2)_{i \in [K]}\right)$ and

$$\tilde{\mu}_{s,t,i} = \tilde{\sigma}_{s,t,i}^2 \left(\frac{\mu_t}{\sigma_0^2} + \frac{B_{s,t,i}}{\sigma^2} \right) , \quad (4)$$
$$\tilde{\sigma}_{s,t,i}^{-2} = \frac{1}{\sigma_0^2} + \frac{N_{s,t,i}}{\sigma^2} .$$

Note that the above is a Gaussian posterior with prior $\mathcal{N}(\mu_t, \sigma_0^2 I_K)$ and $N_{s,t,i}$ observations.

4.2 Linear Bandit with Gaussian Rewards

Now we study a d-dimensional linear bandit, which is instantiated as (2) as follows. The task parameter $\theta_{s,*}$ are coefficients in a linear model. The covariance matrices Σ_q are Σ_0 are positive semi-definite and known. The reward distribution of action a is $\mathcal{N}(a^\top \theta_{s,*}, \sigma^2)$, where $\sigma > 0$ is a known reward noise.

Similarly to Section 4.1, we obtain closed-form posteriors using Kveton et al. (2021). The hyper-posterior in round t is $Q_t = \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t)$, where

$$\begin{aligned}
\bar{\mu}_t &= \bar{\Sigma}_t \Big(\Sigma_q^{-1} \mu_q + \sum_{s \in [m]} B_{s,t} - G_{s,t} (\Sigma_0^{-1} + G_{s,t})^{-1} B_{s,t} \Big) \\
&= \bar{\Sigma}_t \Big(\Sigma_q^{-1} \mu_q + \sum_{s \in [m]} (\Sigma_0 + G_{s,t}^{-1})^{-1} G_{s,t}^{-1} B_{s,t} \Big) , \\
\bar{\Sigma}_t^{-1} &= \Sigma_q^{-1} + \sum_{s \in [m]} G_{s,t} - G_{s,t} (\Sigma_0^{-1} + G_{s,t})^{-1} G_{s,t} \\
&= \Sigma_q^{-1} + \sum_{s \in [m]} (\Sigma_0 + G_{s,t}^{-1})^{-1} . \quad (5)
\end{aligned}$$

Here

$$G_{s,t} = \sigma^{-2} \sum_{\ell < t} \mathbb{1}\{s \in \mathcal{S}_\ell\} A_{s,\ell} A_{s,\ell}^\top$$

is the outer product of the features of taken actions in task s up to round t and

$$B_{s,t} = \sigma^{-2} \sum_{\ell < t} \mathbb{1}\{s \in \mathcal{S}_\ell\} A_{s,\ell} Y_{s,\ell}$$

is their sum weighted by the observed rewards. Similarly to (3), it is helpful to view (5) as a multivariate Gaussian posterior where each task is a single observation. The observation of task s is the least squares estimate of $\theta_{s,*}$ from task s, $G_{s,t}^{-1} B_{s,t}$, and its covariance is $\Sigma_0 + G_{s,t}^{-1}$. Again, the tasks with many observations affect the value of $\bar{\mu}_t$ more, because $G_{s,t}^{-1}$ approaches a zero matrix in these tasks. In this setting, the covariance approaches Σ_0, because even the unknown task parameter $\theta_{s,*}$ would be a noisy observation of μ_* with covariance Σ_0.

After the hyper-parameter is sampled, $\mu_t \sim Q_t$, the task parameter is sampled, $\theta_{s,t} \sim \mathcal{N}(\tilde{\mu}_{s,t}, \tilde{\Sigma}_{s,t})$, where

$$\begin{aligned}
\tilde{\mu}_{s,t} &= \tilde{\Sigma}_{s,t} \left(\Sigma_0^{-1} \mu_t + B_{s,t} \right) , \quad (6) \\
\tilde{\Sigma}_{s,t}^{-1} &= \Sigma_0^{-1} + G_{s,t} .
\end{aligned}$$

The above is the posterior of a linear model with a Gaussian prior $\mathcal{N}(\mu_t, \Sigma_0)$ and Gaussian observations.

5 KEY IDEAS IN OUR ANALYSES

This section reviews key ideas in our regret analyses, including a novel variance decomposition for the posterior of a hierarchical Gaussian model. Due to space

constraints, we only discuss the linear bandit in Section 4.2.

5.1 Bayes Regret Bound

Fix round t and task $s \in \mathcal{S}_t$. Since `HierTS` is a posterior sampling algorithm, both the posterior sample $\theta_{s,t}$ and the unknown task parameter $\theta_{s,*}$ are i.i.d. conditioned on H_t. Moreover, (1) is a marginalization and conditioning in a hierarchical Gaussian model given in Figure 1. Therefore, although we never explicitly derive $\theta_{s,*} \mid H_t$, we know that it is a multivariate Gaussian distribution (Koller and Friedman, 2009); and we denote it by $\mathbb{P}\left(\theta_{s,*} = \theta \mid H_t\right) = \mathcal{N}(\theta; \hat{\mu}_{s,t}, \hat{\Sigma}_{s,t})$.

Following existing Bayes regret analyses (Russo and Van Roy, 2014), we have that

$$\mathbb{E}\left[A_{s,*}^\top \theta_{s,*} - A_{s,t}^\top \theta_{s,*} \mid H_t\right] = \mathbb{E}\left[A_{s,*}^\top (\theta_{s,*} - \hat{\mu}_{s,t}) \mid H_t\right] + \mathbb{E}\left[A_{s,t}^\top (\hat{\mu}_{s,t} - \theta_{s,*}) \mid H_t\right] .$$

Conditioned on history H_t, we observe that $\hat{\mu}_{s,t} - \theta_{s,*}$ is a zero-mean random vector and that $A_{s,t}$ is independent of it. Hence $\mathbb{E}\left[A_{s,t}^\top (\hat{\mu}_{s,t} - \theta_{s,*}) \,\middle|\, H_t\right] = 0$ and the Bayes regret is bounded as

$$\mathcal{BR}(m,n) \leq \mathbb{E}\left[\sum_{t \geq 1} \sum_{s \in \mathcal{S}_t} \mathbb{E}\left[A_{s,*}^\top (\theta_{s,*} - \hat{\mu}_{s,t}) \mid H_t\right]\right] .$$

The following lemma provides an upper bound on the Bayes regret for m tasks, with at most n interactions with each, using the sum of posterior variances

$$\mathcal{V}(m,n) = \mathbb{E}\left[\sum_{t \geq 1} \sum_{s \in \mathcal{S}_t} \|A_{s,t}\|^2_{\hat{\Sigma}_{s,t}}\right] . \tag{7}$$

The proof is deferred to Appendix A.

Lemma 1. *For any $\delta > 0$, the Bayes regret $\mathcal{BR}(m,n)$ in a hierarchical linear bandit (Section 4.2) is bounded by*

$$\sqrt{2dmn\mathcal{V}(m,n)\log(1/\delta)} + \sqrt{2/\pi}\sigma_{\max} d^{\frac{3}{2}} mn\delta ,$$

where $\sigma^2_{\max} = \lambda_1(\Sigma_0) + \lambda_1^2(\Sigma_0)\lambda_1(\Sigma_q)/\lambda_d^2(\Sigma_0)$. When the action space is finite, $|\mathcal{A}| = K$, we also get

$$\sqrt{2mn\mathcal{V}(m,n)\log(1/\delta)} + \sqrt{2/\pi}\sigma_{\max} K mn\delta .$$

Therefore, to bound the regret, we only need to bound the posterior variances induced by the taken actions. The main challenge is that our posterior is over multiple variables. As can be seen in (1), it comprises the hyper-posterior Q_t over μ_* and the conditional $P_{s,t}$ over $\theta_{s,*}$. For any fixed μ_*, $P_{s,t}(\cdot \mid \mu_*)$ should concentrate at $\theta_{s,*}$ as the agent gets more observations from task s. In addition, Q_t should concentrate at μ_* as the agent learns more about μ_* from all tasks.

5.2 Total Variance Decomposition

Due to the hierarchical structure of our problem, it is difficult to reason about the rate at which $\hat{\Sigma}_{s,t}$ "decreases". In this work, we propose a novel variance decomposition that allows this. The decomposition uses the law of total variance (Weiss, 2005), which states that for any X and Y,

$$\mathrm{var}\left[X\right] = \mathbb{E}\left[\mathrm{var}\left[X \mid Y\right]\right] + \mathrm{var}\left[\mathbb{E}\left[X \mid Y\right]\right] .$$

If $X = \theta$ was a scalar task parameter and $Y = \mu$ was a scalar hyper-parameter, and we conditioned on H, the law would give

$$\mathrm{var}\left[\theta \mid H\right] = \mathbb{E}\left[\mathrm{var}\left[\theta \mid \mu, H\right] \mid H\right] + \mathrm{var}\left[\mathbb{E}\left[\theta \mid \mu, H\right] \mid H\right] .$$

This law extends to covariances (Weiss, 2005), where the conditional variance $\mathrm{var}\left[\cdot \mid H\right]$ is substituted with the covariance $\mathrm{cov}\left[\cdot \mid H\right]$. We show the decomposition for a hierarchical Gaussian model below.

5.3 Hierarchical Gaussian Models

Recall that $\hat{\Sigma}_{s,t} = \mathrm{cov}\left[\theta_{s,*} \mid H_t\right]$. We derive a general formula for decomposing $\mathrm{cov}\left[\theta_{s,*} \mid H_t\right]$ below. To simplify notation, we consider a fixed task s and round t, and drop subindexing by them.

Lemma 2. *Let $\theta \mid \mu \sim \mathcal{N}(\mu, \Sigma_0)$ and $H = (x_t, Y_t)_{t=1}^n$ be n observations generated as $Y_t \mid \theta, x_t \sim \mathcal{N}(x_t^\top \theta, \sigma^2)$. Let $\mathbb{P}\left(\mu \mid H\right) = \mathcal{N}(\mu; \bar{\mu}, \bar{\Sigma})$. Then*

$$\mathrm{cov}\left[\theta \mid H\right] = (\Sigma_0^{-1} + G)^{-1} + (\Sigma_0^{-1} + G)^{-1}\Sigma_0^{-1}\bar{\Sigma}\Sigma_0^{-1}(\Sigma_0^{-1} + G)^{-1} ,$$

where $G = \sigma^{-2}\sum_{t=1}^n x_t x_t^\top$. Moreover, for any $x \in \mathbb{R}^d$,

$$x^\top (\Sigma_0^{-1} + G)^{-1}\Sigma_0^{-1}\bar{\Sigma}\Sigma_0^{-1}(\Sigma_0^{-1} + G)^{-1} x \leq \frac{\lambda_1^2(\Sigma_0)\lambda_1(\bar{\Sigma})}{\lambda_d^2(\Sigma_0)}\|x\|_2^2 .$$

Proof. By definition,

$$\mathrm{cov}\left[\theta \mid \mu, H\right] = (\Sigma_0^{-1} + G)^{-1} ,$$
$$\mathbb{E}\left[\theta \mid \mu, H\right] = \mathrm{cov}\left[\theta \mid \mu, H\right](\Sigma_0^{-1}\mu + B) ,$$

where $B = \sigma^{-2}\sum_{t=1}^n x_t Y_t$. Because $\mathrm{cov}\left[\theta \mid \mu, H\right]$ does not depend on μ, $\mathbb{E}\left[\mathrm{cov}\left[\theta \mid \mu, H\right] \mid H\right] = \mathrm{cov}\left[\theta \mid \mu, H\right]$. In addition, since B is a constant conditioned on H,

$$\begin{aligned}\mathrm{cov}&\left[\mathbb{E}\left[\theta \mid \mu, H\right] \mid H\right] \\ &= \mathrm{cov}\left[\mathrm{cov}\left[\theta \mid \mu, H\right]\Sigma_0^{-1}\mu \,\middle|\, H\right] \\ &= (\Sigma_0^{-1} + G)^{-1}\Sigma_0^{-1}\bar{\Sigma}\Sigma_0^{-1}(\Sigma_0^{-1} + G)^{-1} .\end{aligned}$$

This proves the first claim. The second claim follows from standard norm and eigenvalue inequalities. □

We use Lemma 2 as follows. For task s and round t, the posterior covariance decomposes as

$$\hat{\Sigma}_{s,t} = (\Sigma_0^{-1} + G_{s,t})^{-1} + (\Sigma_0^{-1} + G_{s,t})^{-1}\Sigma_0^{-1}\bar{\Sigma}_t\Sigma_0^{-1}(\Sigma_0^{-1} + G_{s,t})^{-1}\,. \tag{8}$$

The first term is $\mathrm{cov}\left[\theta_{s,*} \,|\, \mu_*, H_t\right]$ and captures uncertainty in $\theta_{s,*}$ conditioned on μ_*. The second term depends on hyper-posterior covariance $\bar{\Sigma}_t$ and represents uncertainty in μ_*. Since the first term is exactly $\tilde{\Sigma}_{s,t}$ in (6), while the second term is weighted by it, both are small when we get enough observations for task s. The above also says that $\|A_{s,t}\|^2_{\hat{\Sigma}_{s,t}} = A_{s,t}^\top \hat{\Sigma}_{s,t} A_{s,t}$ in (7) decompose into the two respective norms, which yields our regret decomposition.

5.4 Extensions

So far, we only focused on hierarchical Gaussian models with known hyper-prior and task prior covariances. This is only because they have closed-form posteriors that are easy to interpret and manipulate, without resorting to approximations (Doucet et al., 2001). This choice simplifies algebra and allows us to focus on the key hierarchical structure of our problem. We believe that the tools developed in this section can be applied more broadly. We discuss this next.

Lemma 1 decomposes the Bayes regret into posterior variances and upper bounds on the regret due to tail events. The posterior variance can be derived for any exponential-family posterior with a conjugate prior. On the other hand, the tail inequalities require sub-Gaussianity, which is a property of many exponential-family distributions.

Lemma 2 decomposes the posterior covariance in a hierarchical Gaussian model. It relies on the law of total covariance, which holds for any distribution, to obtain the task and hyper-parameter uncertainties. We expect that similar lemmas can be proved for other hierarchical models, so long as closed-form expressions for the respective uncertainties exist. Another notable property of our decomposition is that it does not require the marginal posterior of $\theta_{s,*}$. We view it as a strength. It means that our approach can be applied to complex graphical models where the marginal uncertainty may be hard to express, but the conditional and prior uncertainties are readily available.

One limitation of our analyses is that we bound the Bayes regret, instead of a stronger frequentist regret. This simplifies our proofs while they still capture our problem structure. Our analyses can be extended to the frequentist setting. This only requires a new proof of Lemma 1, with martingale bounds for tail events and anti-concentration bounds for posterior sampling. The rest of the analysis, where our main contributions are, would not change.

6 REGRET BOUNDS

This section bounds the Bayes regret of `HierTS` in the linear bandit in Section 4.2. Our bounds are specialized to multi-armed bandits in Appendix D. The key idea is to bound the posterior variances $\mathcal{V}(m,n)$ in (7) and then substitute the bound into the infinite-action bound in Lemma 1. We bound the variances using the total covariance decomposition in Section 5.2. Without loss of generality, we assume that the action set $\mathcal{A}$ is a subset of a unit ball, that is $\max_{a\in\mathcal{A}} \|a\|_2 \leq 1$ for any action $a \in \mathcal{A}$.

We make the following contributions in theory. First, we prove regret bounds using a novel variance decomposition (Section 5.2), which improves in constant factors over classical information-theory bounds (Russo and Van Roy, 2016). Second, we prove the first Bayes regret bound for the setting where an agent that interacts with multiple tasks simultaneously.

This section has two parts. In Section 6.1, we assume that only one action is taken in any round t, $|\mathcal{S}_t| = 1$. We call this setting *sequential*, and note that it is the primary setting studied by prior works (Kveton et al., 2021; Basu et al., 2021). In Section 6.3, we focus on a *concurrent* setting, where a single action can be taken in up to L tasks in any round t, $|\mathcal{S}_t| \leq L \leq m$. The challenge of this setting is that the task parameters are only updated after all actions are taken.

6.1 Sequential Regret

The following theorem provides a regret bound for the sequential setting.

Theorem 3 (Sequential regret)**.** *Let $|\mathcal{S}_t| = 1$ for all rounds t and $\mathcal{A} \subseteq \mathbb{R}^d$. Choose $\delta = 1/(mn)$. Then the Bayes regret of* `HierTS` *is*

$$\mathcal{BR}(m,n) \leq d\sqrt{2mn[c_1 m + c_2]\log(mn)} + c_3\,,$$

where $c_3 = O(d^{\frac{3}{2}})$,

$$c_1 = \frac{\lambda_1(\Sigma_0)}{\log(1+\sigma^{-2}\lambda_1(\Sigma_0))}\log\left(1 + \frac{\lambda_1(\Sigma_0)n}{\sigma^2 d}\right)\,,$$

$$c_2 = \frac{c_q c}{\log(1+\sigma^{-2}c_q)}\log\left(1 + \frac{\lambda_1(\Sigma_q)m}{\lambda_d(\Sigma_0)}\right)\,,$$

$$c_q = \frac{\lambda_1^2(\Sigma_0)\lambda_1(\Sigma_q)}{\lambda_d^2(\Sigma_0)}, \quad c = 1 + \sigma^{-2}\lambda_1(\Sigma_0)\,.$$

The proof of Theorem 3 is based on three steps. First, we use Lemma 1. Second, we employ Lemma 2 to decompose the posterior variance in any round into that

of the task parameters and hyper-parameter. Finally, we apply elliptical lemmas to bound each term separately. Theorem 3 has a nice interpretation: $m\sqrt{c_1 n}$ is the regret for learning task parameters and $\sqrt{c_2 mn}$ is the regret for learning the hyper-parameter μ_*. We elaborate on both terms below.

The term $m\sqrt{c_1 n}$ represents the regret for solving m bandit tasks, which are sampled i.i.d. from a known prior $\mathcal{N}(\mu_*, \Sigma_0)$. Under this assumption, no task provides information about any other task, and thus the term is linear in m. The constant c_1 is $O(\lambda_1(\Sigma_0))$ and reflects the dependence on the prior width $\sqrt{\lambda_1(\Sigma_0)}$. Roughly speaking, when the task prior is half as informative, $\sqrt{c_1}$ doubles and so does $m\sqrt{c_1 n}$. This is the expected scaling with conditional uncertainty of $\theta_{s,*}$ given μ_*.

The term $\sqrt{c_2 mn}$ is the regret for learning the hyper-parameter. Asymptotically, it is $O(\sqrt{m})$ smaller than $m\sqrt{c_1 n}$. Therefore, for a large number of tasks m, its contribution to the total regret is negligible. This is why hierarchical Bayesian bandits perform so well in practice. The constant c_2 is $O(\lambda_1(\Sigma_q))$ and reflects the dependence on the hyper-prior width $\sqrt{\lambda_1(\Sigma_0)}$. When the hyper-prior is half as informative, $\sqrt{c_2}$ doubles and so does $\sqrt{c_2 mn}$. This is the expected scaling with the marginal uncertainty of μ_*.

6.2 Tightness of Regret Bounds

One shortcoming of our current analysis is that we do not provide a matching lower bound. To the best of our knowledge, Bayes regret lower bounds are rare and do not match existing upper bounds. The only lower bound that we are aware of is $\Omega(\log^2 n)$ in Theorem 3 of Lai (1987). The bound is for K-armed bandits and it is unclear how to apply it to structured problems. Seminal works on Bayes regret minimization (Russo and Van Roy, 2014, 2016) do not match it. Therefore, to show that our problem structure is reflected in our bound, we compare the regret of `HierTS` to baselines that have more information or use less structure.

Now we compare the regret of `HierTS` to two `LinTS` (Agrawal and Goyal, 2013) baselines that do use our hierarchical model. This first is an oracle `LinTS` that knows μ_*, and so has more information than `HierTS`. Its Bayes regret would be as in Theorem 3 with $c_2 = 0$. Not surprisingly, it is lower than that of `HierTS`. The second baseline is `LinTS` that knows that $\mu_* \sim \mathcal{N}(\mu_q, \Sigma_q)$, but does not model the structure that the tasks share μ_*. In this case, each task parameter can be viewed as having prior $\mathcal{N}(\mu_q, \Sigma_q + \Sigma_0)$. The regret of this algorithm would be as in Theorem 3 with $c_2 = 0$, while $\lambda_1(\Sigma_0)$ in c_1 would be $\lambda_1(\Sigma_q + \Sigma_0)$. Since c_1 is multiplied by m while c_2 is not, `HierTS` would have lower regret as $m \to \infty$. This is a powerful testament to the benefit of learning the hyper-parameter.

Finally, we want to comment on linear dependence in d and m in Theorem 3. The dependence on d is standard in Bayes regret analyses for linear bandits with infinitely many arms (Russo and Van Roy, 2014; Lu and Van Roy, 2019). As for the number of tasks m, since the tasks are drawn i.i.d. from the same hyper-prior, they do not provide any additional information about each other. So, even if the hyper-parameter is known, the regret for learning to act in m tasks with n rounds would be $O(m\sqrt{n})$. Our improvements are in constants due to better variance attribution. Other bandit meta-learning works (Kveton et al., 2021; Basu et al., 2021) made similar observations. Also note that the frequentist regret of `LinTS` applied to m independent linear bandit tasks is $\tilde{O}(md^{\frac{3}{2}}\sqrt{n})$ (Agrawal and Goyal, 2013). This is worse by a factor of $\sqrt{d}$ than the bound in Theorem 3.

6.3 Concurrent Regret

Now we investigate the concurrent setting, where the agent acts in up to L tasks per round. This setting is challenging because the hyper-posterior Q_t is not updated until the end of the round. This is because the task posteriors are not refined with the observations from concurrent tasks. This delayed feedback should increase regret. Before we show it, we make the following assumption on the action space.

Assumption 1. *There exist actions $\{a_i\}_{i=1}^d \subseteq \mathcal{A}$ and $\eta > 0$ such that $\lambda_d(\sum_{i=1}^d a_i a_i^\top) \geq \eta$.*

This assumption is without loss of generality. Specifically, if $\mathbb{R}^d$ was not spanned by actions in $\mathcal{A}$, we could project $\mathcal{A}$ into a subspace where the assumption would hold. Our regret bound is below.

Theorem 4 (Concurrent regret)**.** *Let $|\mathcal{S}_t| \leq L \leq m$ and $\mathcal{A} \subseteq \mathbb{R}^d$. Let $\delta = 1/(mn)$. Then the Bayes regret of* `HierTS` *is*

$$\mathcal{BR}(m, n) \leq d\sqrt{2mn[c_1 m + c_2]\log(mn)} + c_3\,,$$

where c_1, c_q, and c are defined as in Theorem 3,

$$c_2 = \frac{c_q c_4 c}{\log(1 + \sigma^{-2} c_q)} \log\left(1 + \frac{\lambda_1(\Sigma_q) m}{\lambda_d(\Sigma_0)}\right)\,,$$
$$c_4 = 1 + \frac{\sigma^{-2}\lambda_1(\Sigma_q)(\lambda_1(\Sigma_0) + \sigma^2/\eta)}{\lambda_1(\Sigma_q) + (\lambda_1(\Sigma_0) + \sigma^2/\eta)/L}\,,$$

and $c_3 = O(d^{\frac{3}{2}} m)$.

The key step in the proof is to modify `HierTS` as follows. For the first d interactions with any task s, we take actions $\{a_i\}_{i=1}^d$. This guarantees that we explore

all directions within the task, and allows us to bound losses from not updating the task posterior with concurrent observations. This modification of `HierTS` is trivial and analogous to popular initialization in bandits, where each arm is pulled once in the first rounds (Auer et al., 2002).

The regret bound in Theorem 4 is similar to that in Theorem 3. There are two key differences. First, the additional scaling factor c_4 in c_2 is the price for taking concurrent actions. It increases as more actions L are taken concurrently, but is sublinear in L. Second, c_3 arises due to trivially bounding dm rounds of forced exploration. To the best of our knowledge, Theorem 4 is the first Bayes regret bound where multiple bandit tasks are solved concurrently. Prior works only proved frequentist regret bounds (Yang et al., 2021).

7 EXPERIMENTS

We compare `HierTS` to two TS baselines (Section 6.2) that do not learn the hyper-parameter μ_*. The first baseline is an idealized algorithm that knows μ_* and uses the true prior $\mathcal{N}(\mu_*, \Sigma_0)$. We call it `OracleTS`. As `OracleTS` has more information than `HierTS`, we expect it to outperform `HierTS`. The second baseline, which we call `TS`, ignores that μ_* is shared among the tasks and uses the marginal prior of $\theta_{s,*}$, $\mathcal{N}(\mu_q, \Sigma_q + \Sigma_0)$, in each task.

We experiment with two linear bandit problems with $m = 10$ tasks: a synthetic problem with Gaussian rewards and an online image classification problem. The former is used to validate our regret bounds. The latter has non-Gaussian rewards and demonstrates that `HierTS` is robust to prior misspecification. Our setup closely follows Basu et al. (2021). However, our tasks can arrive in an arbitrary order and in parallel. Due to space constraints, we only report the synthetic experiment here, and defer the rest to Appendix E.

The synthethic problem is defined as follows: $d = 2$, $|\mathcal{A}| = 10$, and each action is sampled uniformly from $[-0.5, 0.5]^d$. Initially, the number of concurrent tasks is $L = 5$; but we vary it later to measure its impact on regret. The number of rounds is $n = 200m/L$ and $\mathcal{S}_t$ is defined as follows. First, we take a random permutation of the list of tasks where each task appears exactly 200 times. Then we batch every L consecutive elements of the list and set $\mathcal{S}_t$ to the t-th batch. The hyper-prior is $\mathcal{N}(\mathbf{0}, \Sigma_q)$ with $\Sigma_q = \sigma_q^2 I_d$, the task covariance is $\Sigma_0 = \sigma_0^2 I_d$, and the reward noise is $\sigma = 0.5$. We choose $\sigma_q \in \{0.5, 1\}$ and $\sigma_0 = 0.1$, where $\sigma_q \gg \sigma_0$ so that the effect of learning μ_* on faster learning of $\theta_{s,*}$ is easier to measure.

The regret of all compared algorithms is reported in Figure 2. In plots (a) and (b), we show how the regret scales with the number of rounds for small ($\sigma_q = 0.5$) and large ($\sigma_q = 1$) hyper-prior width. As suggested in Section 6.1, `HierTS` outperforms `TS` that does not try to learn μ_*. It is comparable to `OracleTS` when σ_q is small, but degrades as σ_q increases. This matches the regret bound in Theorem 4, where c_2 grows with σ_q. In plot (c), we show how the regret of `HierTS` varies with the number of concurrent tasks L. We observe that it increases with L, but the increase is sublinear, as suggested in Section 6.3.

8 RELATED WORK

The most related works are recent papers on bandit meta-learning (Bastani et al., 2019; Ortega et al., 2019; Cella et al., 2020; Kveton et al., 2021; Basu et al., 2021; Peleg et al., 2021; Simchowitz et al., 2021), where a learning agent interacts with a single task at a time until completion. Both Kveton et al. (2021) and Basu et al. (2021) represent their problems using graphical models and apply Thompson sampling to solve them. The setting of these papers is less general than ours. Wan et al. (2021) study a setting where the tasks can arrive in any order. We differ from this work in several aspects. First, they only consider a K-armed bandit. Second, their model is different. In our notation, Wan et al. (2021) assume that the mean reward of action a in task s is $x_{s,a}^\top \mu_*$ plus i.i.d. noise, where $x_{s,a}$ is an observed feature vector. The i.i.d. noise prevents generalization to a large number of actions. In our work, the mean reward of action a in task s is $a^\top \theta_{s,*}$, where $\theta_{s,*} \sim \mathcal{N}(\mu_*, \Sigma_0)$. Third, Wan et al. (2021) derive a frequentist regret bound, which matches Theorem 3 asymptotically, but does not explicitly depend on prior widths. Finally, Wan et al. (2021) do not consider the concurrent setting. To the best of our knowledge, we are the first to study Bayesian bandits with arbitrarily ordered and concurrent tasks.

The novelty in our analysis is the total covariance decomposition, which leads to better variance attribution in structured models than information-theoretic bounds (Russo and Van Roy, 2016; Lu and Van Roy, 2019; Basu et al., 2021). For instance, take Theorem 5 of Basu et al. (2021), which corresponds to our sequential meta-learning setting. Forced exploration is needed to make their task term $O(\lambda_1(\Sigma_0))$. This is because the upper bound on the regret with filtered mutual information depends on the maximum marginal task parameter covariance, which can be $\lambda_1(\Sigma_q + \Sigma_0)$. In our analysis, the comparable term c_1 (Theorem 3) is $O(\lambda_1(\Sigma_0))$ without any forced exploration. We also improve upon related analysis of Kveton et al. (2021) in several aspects. First, Kveton et al. (2021) analyze only a K-armed bandit. Second, they derive that the

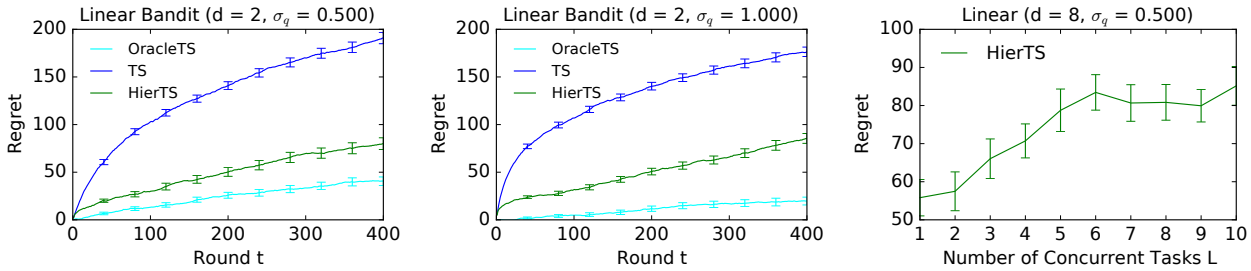

Figure 2: Evaluation of `HierTS` on synthetic bandit problems. From left to right, we report the Bayes regret (a) for smaller σ_q, (b) for larger σ_q, (c) and as a function of the number of concurrent tasks L.

additional regret for meta-learning is $\tilde{O}(\sqrt{m}n^2)$; while our bound shows $\tilde{O}(\sqrt{mn})$. Finally, our setting generalizes bandit meta-learning.

Meta- and multi-task bandits have also been studied in the frequentist setting (Azar et al., 2013; Deshmukh et al., 2017). Cella et al. (2020) propose a `LinUCB` algorithm (Abbasi-Yadkori et al., 2011) that constructs an ellipsoid around the unknown hyper-parameter in a linear bandit. The concurrent setting has also been studied, but with a different shared structure of task parameters. Dubey and Pentland (2020) use a kernel matrix, Wang et al. (2021) utilize pairwise distances of task parameters, and Yang et al. (2021) use low-rank factorization. Our structure, where the task parameters are drawn from an unknown prior, is both novel and important to study because it differs significantly from the aforementioned works. Earlier works on bandits with similar instances rely on clustering (Gentile et al., 2014, 2017; Li et al., 2016) and low-rank factorization (Kawale et al., 2015; Sen et al., 2017; Katariya et al., 2016, 2017). They analyze the frequentist regret, which is a stronger metric than the Bayes regret. Except for one work, all algorithms are UCB-like and conservative in practice. In comparison, `HierTS` uses a natural stochastic structure. This makes it practical, to the point that the analyzed algorithm performs well in practice without any additional tuning.

Another related line of work are latent bandits (Maillard and Mannor, 2014; Hong et al., 2020, 2022), where the bandit problem is parameterized by an unknown latent state. If known, the latent state could help the agent to identify the bandit instance that it interacts with. These works reason about latent variables; but the purpose is different from our work, where we introduce the unknown hyper-parameter μ_* to relate multiple similar tasks.

9 CONCLUSIONS

We study *hierarchical Bayesian bandits*, a general setting for solving similar bandit tasks. Instances of our setting recover meta-, multi-task, and federated bandits in prior works. We propose a natural hierarchical Thompson sampling algorithm, which can be implemented exactly and analyzed in Gaussian models. We analyze it using a novel total variance decomposition, which leads to interpretable regret bounds that scale with the hyper-prior and task prior widths. The benefit of hierarchical models is shown in both synthetic and real-world domains.

While we view our work as solving an extremely general problem, there are multiple directions for future work. For instance, we only study a specific hierarchical Gaussian structure in Section 4. However, based on the discussion in Section 5.4, we believe that our tools would apply to arbitrary graphical models with general sub-Gaussian distributions. Another direction for future work are frequentist upper bounds and matching lower bounds, in both the frequentist and Bayesian settings.

References

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

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceeding of the 25th Annual Conference on Learning Theory*, pages 39.1–39.26, 2012.

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

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

M. G. Azar, A. Lazaric, and E. Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In *Advances in Neural Information Processing Systems 26*, pages 2220–2228, 2013.

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

S. Basu, B. Kveton, M. Zaheer, and C. Szepesvari. No regrets for learning the prior in bandits. In *Advances in Neural Information Processing Systems 34*, 2021.

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

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

W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.

V. Dani, T. Hayes, and S. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 355–366, 2008.

A. A. Deshmukh, U. Dogan, and C. Scott. Multi-task learning for contextual bandits. In *Advances in Neural Information Processing Systems 30*, pages 4848–4856, 2017.

A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, New York, NY, 2001.

A. Dubey and A. Pentland. Kernel methods for cooperative multi-agent contextual bandits. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Y. Gai, B. Krishnamachari, and R. Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

A. Garivier and O. Cappe. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceeding of the 24th Annual Conference on Learning Theory*, pages 359–376, 2011.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2013.

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

C. Gentile, S. Li, P. Kar, A. Karatzoglou, G. Zappella, and E. Etrue. On context-dependent clustering of bandits. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

J. Hong, B. Kveton, M. Zaheer, Y. Chow, A. Ahmed, and C. Boutilier. Latent bandits revisited. In *Advances in Neural Information Processing Systems 33*, 2020.

J. Hong, B. Kveton, M. Zaheer, M. Ghavamzadeh, and C. Boutilier. Thompson sampling with a mixture prior. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

S. Katariya, B. Kveton, C. Szepesvari, and Z. Wen. DCM bandits: Learning to rank with multiple clicks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1215–1224, 2016.

S. Katariya, B. Kveton, C. Szepesvari, C. Vernade, and Z. Wen. Stochastic rank-1 bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

J. Kawale, H. Bui, B. Kveton, L. Tran-Thanh, and S. Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems 28*, pages 1297–1305, 2015.

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

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015a.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015b.

B. Kveton, M. Konobeev, M. Zaheer, C.-W. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

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

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

T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

Y. LeCun, C. Cortes, and C. Burges. MNIST Handwritten Digit Database. http://yann.lecun.com/exdb/mnist, 2010.

S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proceedings of the 39th Annual International ACM SIGIR Conference*, 2016.

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

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

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

A. Moradipari, B. Turan, Y. Abbasi-Yadkori, M. Alizadeh, and M. Ghavamzadeh. Parameter and feature selection in stochastic linear bandits. *CoRR*, abs/2106.05378, 2021. URL https://arxiv.org/abs/2106.05378.

P. Ortega, J. Wang, M. Rowland, T. Genewein, Z. Kurth-Nelson, R. Pascanu, N. Heess, J. Veness, A. Pritzel, P. Sprechmann, S. Jayakumar, T. McGrath, K. Miller, M. G. Azar, I. Osband, N. Rabinowitz, A. Gyorgy, S. Chiappa, S. Osindero, Y. W. Teh, H. van Hasselt, N. de Freitas, M. Botvinick, and S. Legg. Meta-learning of sequential strategies. *CoRR*, abs/1905.03030, 2019. URL http://arxiv.org/abs/1905.03030.

A. Peleg, N. Pearl, and R. Meir. Metalearning linear bandits by prior update. *CoRR*, abs/2107.05320, 2021. URL https://arxiv.org/abs/2107.05320.

F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pages 784–791, 2008.

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

D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

R. Sen, K. Shanmugam, M. Kocaoglu, A. Dimakis, and S. Shakkottai. Contextual bandits with latent confounders: An NMF approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

C. Shi and C. Shen. Federated multi-armed bandits. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

M. Simchowitz, C. Tosh, A. Krishnamurthy, D. Hsu, T. Lykouris, M. Dudik, and R. Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. In *Advances in Neural Information Processing Systems 34*, 2021.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

R. Wan, L. Ge, and R. Song. Metadata-based multitask bandits with bayesian hierarchical models. In *Advances in Neural Information Processing Systems 34*, 2021.

Z. Wang, C. Zhang, M. K. Singh, L. Riek, and K. Chaudhuri. Multitask bandit learning through heterogeneous feedback aggregation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

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

J. Yang, W. Hu, J. Lee, and S. Du. Impact of representation learning in bandits. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Y. Zhang and Q. Yang. A survey on multi-task learning. *CoRR*, abs/1707.08114, 2017. URL http://arxiv.org/abs/1707.08114.

A Proof of Lemma 1

The first claim is proved as follows. Fix round t and task $s \in \mathcal{S}_t$. Since $\hat{\mu}_{s,t}$ is a deterministic function of H_t, and $A_{s,*}$ and $A_{s,t}$ are i.i.d. given H_t, we have

$$\mathbb{E}\left[A_{s,*}^\top \theta_{s,*} - A_{s,t}^\top \theta_{s,*}\right] = \mathbb{E}\left[\mathbb{E}\left[A_{s,*}^\top(\theta_{s,*} - \hat{\mu}_{s,t}) \,\middle|\, H_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[A_{s,t}^\top(\hat{\mu}_{s,t} - \theta_{s,*}) \,\middle|\, H_t\right]\right].$$

Moreover, $\theta_{s,*} - \hat{\mu}_{s,t}$ is a zero-mean random vector independent of $A_{s,t}$, and thus $\mathbb{E}\left[A_{s,t}^\top(\hat{\mu}_{s,t} - \theta_{s,*}) \,\middle|\, H_t\right] = 0$. So we only need to bound the first term above. Let

$$E_{s,t} = \left\{\|\theta_{s,*} - \hat{\mu}_{s,t}\|_{\hat{\Sigma}_{s,t}^{-1}} \le \sqrt{2d\log(1/\delta)}\right\}$$

be the event that a high-probability confidence interval for the task parameter $\theta_{s,*}$ holds. Fix history H_t. Then by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}\left[A_{s,*}^\top(\theta_{s,*} - \hat{\mu}_{s,t}) \,\middle|\, H_t\right] &\le \mathbb{E}\left[\|A_{s,*}\|_{\hat{\Sigma}_{s,t}} \|\theta_{s,*} - \hat{\mu}_{s,t}\|_{\hat{\Sigma}_{s,t}^{-1}} \,\middle|\, H_t\right] \\
&\le \sqrt{2d\log(1/\delta)}\,\mathbb{E}\left[\|A_{s,*}\|_{\hat{\Sigma}_{s,t}} \,\middle|\, H_t\right] + \underbrace{\max_{a\in\mathcal{A}} \|a\|_{\hat{\Sigma}_{s,t}}}_{\le \sigma_{\max}} \mathbb{E}\left[\|\theta_{s,*} - \hat{\mu}_{s,t}\|_{\hat{\Sigma}_{s,t}^{-1}} \mathbb{1}\{\bar{E}_{s,t}\} \,\middle|\, H_t\right] \\
&= \sqrt{2d\log(1/\delta)}\,\mathbb{E}\left[\|A_{s,t}\|_{\hat{\Sigma}_{s,t}} \,\middle|\, H_t\right] + \sigma_{\max}\,\mathbb{E}\left[\|\theta_{s,*} - \hat{\mu}_{s,t}\|_{\hat{\Sigma}_{s,t}^{-1}} \mathbb{1}\{\bar{E}_{s,t}\} \,\middle|\, H_t\right].
\end{aligned}$$

The equality follows from the fact that $\hat{\Sigma}_{s,t}$ is a deterministic function of H_t, and that $A_{s,*}$ and $A_{s,t}$ are i.i.d. given H_t. Now we focus on the second term above. First, note that

$$\|\theta_{s,*} - \hat{\mu}_{s,t}\|_{\hat{\Sigma}_{s,t}^{-1}} = \|\hat{\Sigma}_{s,t}^{-\frac{1}{2}}(\theta_{s,*} - \hat{\mu}_{s,t})\|_2 \le \sqrt{d}\|\hat{\Sigma}_{s,t}^{-\frac{1}{2}}(\theta_{s,*} - \hat{\mu}_{s,t})\|_\infty.$$

By definition, $\theta_{s,*} - \hat{\mu}_{s,t} \mid H_t \sim \mathcal{N}(\mathbf{0}, \hat{\Sigma}_{s,t})$, and hence $\hat{\Sigma}_{s,t}^{-\frac{1}{2}}(\theta_{s,*} - \hat{\mu}_{s,t}) \mid H_t$ is a d-dimensional standard normal variable. Moreover, note that $\bar{E}_{s,t}$ implies $\|\hat{\Sigma}_{s,t}^{-\frac{1}{2}}(\theta_{s,*} - \hat{\mu}_{s,t})\|_\infty \ge \sqrt{2\log(1/\delta)}$. Finally, we combine these facts with a union bound over all entries of $\hat{\Sigma}_{s,t}^{-\frac{1}{2}}(\theta_{s,*} - \hat{\mu}_{s,t}) \mid H_t$, which are standard normal variables, and get

$$\mathbb{E}\left[\|\hat{\Sigma}_{s,t}^{-\frac{1}{2}}(\theta_{s,*} - \hat{\mu}_{s,t})\|_\infty \mathbb{1}\{\bar{E}_{s,t}\} \,\middle|\, H_t\right] \le 2\sum_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \int_{u=\sqrt{2\log(1/\delta)}}^{\infty} u \exp\left[-\frac{u^2}{2}\right] \mathrm{d}u \le \sqrt{\frac{2}{\pi}} d\delta.$$

Now we combine all inequalities and have

$$\mathbb{E}\left[A_{s,*}^\top(\theta_{s,*} - \hat{\mu}_{s,t}) \,\middle|\, H_t\right] \le \sqrt{2d\log(1/\delta)}\,\mathbb{E}\left[\|A_{s,t}\|_{\hat{\Sigma}_{s,t}} \,\middle|\, H_t\right] + \sqrt{\frac{2}{\pi}}\sigma_{\max} d^{\frac{3}{2}}\delta.$$

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

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\ge1}\sum_{s\in\mathcal{S}_t} A_{s,*}^\top\theta_{s,*} - A_{s,t}^\top\theta_{s,*}\right] &\le \sqrt{2d\log(1/\delta)}\,\mathbb{E}\left[\sum_{t\ge1}\sum_{s\in\mathcal{S}_t}\|A_{s,t}\|_{\hat{\Sigma}_{s,t}}\right] + \sqrt{\frac{2}{\pi}}\sigma_{\max} d^{\frac{3}{2}} mn\delta \\
&\le \sqrt{2dmn\log(1/\delta)}\sqrt{\mathbb{E}\left[\sum_{t\ge1}\sum_{s\in\mathcal{S}_t}\|A_{s,t}\|_{\hat{\Sigma}_{s,t}}^2\right]} + \sqrt{\frac{2}{\pi}}\sigma_{\max} d^{\frac{3}{2}} mn\delta.
\end{aligned}$$

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

To bound $\sigma_{\max}$, we use Weyl's inequalities together with (8), the second claim in Lemma 2, and (5). Specifically, under the assumption that $\|a\|_2 \le 1$ for all $a \in \mathcal{A}$, we have

$$\begin{aligned}
\max_{a\in\mathcal{A}}\|a\|_{\hat{\Sigma}_{s,t}}^2 &\le \lambda_1(\hat{\Sigma}_{s,t}) \le \lambda_1((\Sigma_0^{-1} + G_{s,t})^{-1}) + \lambda_1((\Sigma_0^{-1} + G_{s,t})^{-1}\Sigma_0^{-1}\bar{\Sigma}_t\Sigma_0^{-1}(\Sigma_0^{-1} + G_{s,t})^{-1}) \\
&\le \lambda_1(\Sigma_0) + \frac{\lambda_1^2(\Sigma_0)\lambda_1(\Sigma_q)}{\lambda_d^2(\Sigma_0)} = \sigma_{\max}^2.
\end{aligned}$$

This concludes the proof of the first claim.

The second claim is proved by modifying the first proof as follows. Fix round t and task $s \in \mathcal{S}_t$. Let

$$E_{s,t} = \left\{\forall a \in \mathcal{A} : |a^\top(\theta_{s,*} - \hat{\mu}_{s,t})| \leq \sqrt{2\log(1/\delta)}\|a\|_{\hat{\Sigma}_{s,t}}\right\}$$

be the event that all high-probability confidence intervals hold. Then we have

$$\mathbb{E}\left[A_{s,*}^\top(\theta_{s,*} - \hat{\mu}_{s,t}) \,\middle|\, H_t\right] \leq \sqrt{2\log(1/\delta)}\,\mathbb{E}\left[\|A_{s,t}\|_{\hat{\Sigma}_{s,t}} \,\middle|\, H_t\right] + \mathbb{E}\left[A_{s,*}^\top(\theta_{s,*} - \hat{\mu}_{s,t})\mathbb{1}\{\bar{E}_{s,t}\} \,\middle|\, H_t\right] .$$

Now note that for any action a, $a^\top(\theta_{s,*} - \hat{\mu}_{s,t})/\|a\|_{\hat{\Sigma}_{s,t}}$ is a standard normal variable. It follows that

$$\mathbb{E}\left[A_{s,*}^\top(\theta_{s,*} - \hat{\mu}_{s,t})\mathbb{1}\{\bar{E}_{s,t}\} \,\middle|\, H_t\right] \leq 2\sum_{a \in \mathcal{A}} \|a\|_{\hat{\Sigma}_{s,t}} \frac{1}{\sqrt{2\pi}} \int_{u=\sqrt{2\log(1/\delta)}}^{\infty} u \exp\left[-\frac{u^2}{2}\right] \mathrm{d}u \leq \sqrt{\frac{2}{\pi}}\sigma_{\max} K\delta\,.$$

The rest of the proof proceeds as in the first claim, yielding

$$\mathbb{E}\left[A_{s,*}^\top(\theta_{s,*} - \hat{\mu}_{s,t}) \,\middle|\, H_t\right] \leq \sqrt{2\log(1/\delta)}\,\mathbb{E}\left[\|A_{s,t}\|_{\hat{\Sigma}_{s,t}} \,\middle|\, H_t\right] + \sqrt{\frac{2}{\pi}}\sigma_{\max} K\delta\,.$$

This completes the proof.

B Proof of Theorem 3

Lemma 1 says that the Bayes regret $\mathcal{BR}(m,n)$ can be bounded by bounding the sum of posterior variances $\mathcal{V}(m,n)$. Since $|\mathcal{S}_t| = 1$, we make two simplifications. First, we replace the set of tasks $\mathcal{S}_t$ by a single task $S_t \in [m]$. Second, there are exactly mn rounds.

Fix round t and task $s = S_t$. To reduce clutter, let $M = \Sigma_0^{-1} + G_{s,t}$. By the total covariance decomposition in Lemma 2, we have that

$$\begin{aligned}
\|A_{s,t}\|_{\hat{\Sigma}_{s,t}}^2 &= \sigma^2 \frac{A_{s,t}^\top \hat{\Sigma}_{s,t} A_{s,t}}{\sigma^2} = \sigma^2\left(\sigma^{-2} A_{s,t}^\top \tilde{\Sigma}_{s,t} A_{s,t} + \sigma^{-2} A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_t\Sigma_0^{-1}M^{-1}A_{s,t}\right) \\
&\leq c_1 \log(1 + \sigma^{-2} A_{s,t}^\top \tilde{\Sigma}_{s,t} A_{s,t}) + c_2 \log(1 + \sigma^{-2} A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_t\Sigma_0^{-1}M^{-1}A_{s,t}) \\
&= c_1 \log\det(I_d + \sigma^{-2}\tilde{\Sigma}_{s,t}^{\frac{1}{2}} A_{s,t} A_{s,t}^\top \tilde{\Sigma}_{s,t}^{\frac{1}{2}}) + c_2 \log\det(I_d + \sigma^{-2}\bar{\Sigma}_t^{\frac{1}{2}}\Sigma_0^{-1}M^{-1}A_{s,t}A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_t^{\frac{1}{2}})\,. \qquad (9)
\end{aligned}$$

The logarithmic terms are introduced using

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

which holds for any $x \in [0,u]$. The resulting constants are

$$c_1 = \frac{\lambda_1(\Sigma_0)}{\log(1 + \sigma^{-2}\lambda_1(\Sigma_0))}\,, \quad c_2 = \frac{c_q}{\log(1+\sigma^{-2}c_q)}\,, \quad c_q = \frac{\lambda_1^2(\Sigma_0)\lambda_1(\Sigma_q)}{\lambda_d^2(\Sigma_0)}\,.$$

The derivation of c_1 uses that

$$A_{s,t}^\top \tilde{\Sigma}_{s,t} A_{s,t} = \lambda_1(\tilde{\Sigma}_{s,t}) = \lambda_d^{-1}(\Sigma_0^{-1} + G_{s,t}) \leq \lambda_d^{-1}(\Sigma_0^{-1}) = \lambda_1(\Sigma_0)\,.$$

The derivation of c_2 follows from

$$A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_t\Sigma_0^{-1}M^{-1}A_{s,t} \leq \lambda_1^2(M^{-1})\lambda_1^2(\Sigma_0^{-1})\lambda_1(\bar{\Sigma}_t) \leq \frac{\lambda_1^2(\Sigma_0)\lambda_1(\Sigma_q)}{\lambda_d^2(\Sigma_0)}\,.$$

This is also proved as the second claim in Lemma 2. Now we focus on bounding the logarithmic terms in (9).

B.1 First Term in (9)

This is a per-instance term and can be rewritten as

$$\log\det(I_d + \sigma^{-2}\tilde{\Sigma}_{s,t}^{\frac{1}{2}} A_{s,t} A_{s,t}^\top \tilde{\Sigma}_{s,t}^{\frac{1}{2}}) = \log\det(\tilde{\Sigma}_{s,t}^{-1} + \sigma^{-2} A_{s,t} A_{s,t}^\top) - \log\det(\tilde{\Sigma}_{s,t}^{-1})\,.$$

When we sum over all rounds with task s, we get telescoping and the contribution of this term is at most

$$\begin{aligned}\sum_{t=1}^{mn} \mathbb{1}\{S_t = s\} \log\det(I_d + \sigma^{-2}\tilde{\Sigma}_{s,t}^{\frac{1}{2}} A_{s,t} A_{s,t}^\top \tilde{\Sigma}_{s,t}^{\frac{1}{2}}) &= \log\det(\tilde{\Sigma}_{s,mn+1}^{-1}) - \log\det(\tilde{\Sigma}_{s,1}^{-1}) = \log\det(\Sigma_0^{\frac{1}{2}} \tilde{\Sigma}_{s,mn+1}^{-1} \Sigma_0^{\frac{1}{2}}) \\ &\le d\log\left(\frac{1}{d}\operatorname{tr}(\Sigma_0^{\frac{1}{2}} \tilde{\Sigma}_{s,mn+1}^{-1} \Sigma_0^{\frac{1}{2}})\right) \le d\log\left(1 + \frac{\lambda_1(\Sigma_0) n}{\sigma^2 d}\right),\end{aligned}$$

where we use that task s appears at most n times. Now we sum over all m tasks and get

$$\sum_{t=1}^{mn} \log\det(I_d + \sigma^{-2}\tilde{\Sigma}_{S_t,t}^{\frac{1}{2}} A_{S_t,t} A_{S_t,t}^\top \tilde{\Sigma}_{S_t,t}^{\frac{1}{2}}) \le dm\log\left(1 + \frac{\lambda_1(\Sigma_0) n}{\sigma^2 d}\right).$$

B.2 Second Term in (9)

This is a hyper-parameter term. Before we analyze it, let $v = \sigma^{-1} M^{-\frac{1}{2}} A_{s,t}$ and note that

$$\begin{aligned}\bar{\Sigma}_{t+1}^{-1} - \bar{\Sigma}_t^{-1} &= (\Sigma_0 + (G_{s,t} + \sigma^{-2} A_{s,t} A_{s,t}^\top)^{-1})^{-1} - (\Sigma_0 + G_{s,t}^{-1})^{-1} \\ &= \Sigma_0^{-1} - \Sigma_0^{-1}(M + \sigma^{-2} A_{s,t} A_{s,t}^\top)^{-1} \Sigma_0^{-1} - (\Sigma_0^{-1} - \Sigma_0^{-1} M^{-1} \Sigma_0^{-1}) \\ &= \Sigma_0^{-1}(M^{-1} - (M + \sigma^{-2} A_{s,t} A_{s,t}^\top)^{-1}) \Sigma_0^{-1} \\ &= \Sigma_0^{-1} M^{-\frac{1}{2}} (I_d - (I_d + \sigma^{-2} M^{-\frac{1}{2}} A_{s,t} A_{s,t}^\top M^{-\frac{1}{2}})^{-1}) M^{-\frac{1}{2}} \Sigma_0^{-1} \\ &= \Sigma_0^{-1} M^{-\frac{1}{2}} (I_d - (I_d + vv^\top)^{-1}) M^{-\frac{1}{2}} \Sigma_0^{-1} \\ &= \Sigma_0^{-1} M^{-\frac{1}{2}} \frac{vv^\top}{1 + v^\top v} M^{-\frac{1}{2}} \Sigma_0^{-1} \\ &= \sigma^{-2} \Sigma_0^{-1} M^{-1} \frac{A_{s,t} A_{s,t}^\top}{1 + v^\top v} M^{-1} \Sigma_0^{-1}\,, \end{aligned} \tag{10}$$

where we first use the Woodbury matrix identity and then the Sherman-Morrison formula. Since $\|A_{s,t}\|_2 \le 1$,

$$1 + v^\top v = 1 + \sigma^{-2} A_{s,t}^\top M^{-1} A_{s,t} \le 1 + \sigma^{-2}\lambda_1(\Sigma_0) = c\,.$$

Based on the above derivations, we bound the second logarithmic term in (9) as

$$\begin{aligned}&\log\det(I_d + \sigma^{-2}\bar{\Sigma}_t^{\frac{1}{2}} \Sigma_0^{-1} M^{-1} A_{s,t} A_{s,t}^\top M^{-1} \Sigma_0^{-1} \bar{\Sigma}_t^{\frac{1}{2}}) \\ &\quad\le c\log\det(I_d + \sigma^{-2}\bar{\Sigma}_t^{\frac{1}{2}} \Sigma_0^{-1} M^{-1} A_{s,t} A_{s,t}^\top M^{-1} \Sigma_0^{-1} \bar{\Sigma}_t^{\frac{1}{2}}/c) \\ &\quad= c\left[\log\det(\bar{\Sigma}_t^{-1} + \sigma^{-2}\Sigma_0^{-1} M^{-1} A_{s,t} A_{s,t}^\top M^{-1} \Sigma_0^{-1}/c) - \log\det(\bar{\Sigma}_t^{-1})\right] \\ &\quad\le c\left[\log\det(\bar{\Sigma}_{t+1}^{-1}) - \log\det(\bar{\Sigma}_t^{-1})\right].\end{aligned}$$

The first inequality holds because $\log(1+x) \le c\log(1 + x/c)$ for any $x \ge 0$ and $c \ge 1$. The second inequality follows from the fact that we have a rank-1 update of $\bar{\Sigma}_t^{-1}$. Now we sum over all rounds and get telescoping

$$\begin{aligned}&\sum_{t=1}^{mn} \log\det(I_d + \sigma^{-2}\bar{\Sigma}_t^{\frac{1}{2}} \Sigma_0^{-1}(\Sigma_0^{-1} + G_{S_t,t})^{-1} A_{S_t,t} A_{S_t,t}^\top (\Sigma_0^{-1} + G_{S_t,t})^{-1} \Sigma_0^{-1} \bar{\Sigma}_t^{\frac{1}{2}}) \\ &\quad\le c\left[\log\det(\bar{\Sigma}_{mn+1}^{-1}) - \log\det(\bar{\Sigma}_1^{-1})\right] = c\log\det(\Sigma_q^{\frac{1}{2}} \bar{\Sigma}_{mn+1}^{-1} \Sigma_q^{\frac{1}{2}}) \le cd\log\left(\frac{1}{d}\operatorname{tr}(\Sigma_q^{\frac{1}{2}} \bar{\Sigma}_{mn+1}^{-1} \Sigma_q^{\frac{1}{2}})\right) \\ &\quad\le cd\log(\lambda_1(\Sigma_q^{\frac{1}{2}} \bar{\Sigma}_{mn+1}^{-1} \Sigma_q^{\frac{1}{2}})) \le cd\log\left(1 + \frac{\lambda_1(\Sigma_q) m}{\lambda_d(\Sigma_0)}\right).\end{aligned}$$

Finally, we combine the upper bounds for both logarithmic terms and get

$$\mathcal{V}(m,n) = \mathbb{E}\left[\sum_{t=1}^{mn} \|A_{S_t,t}\|^2_{\hat{\Sigma}_{S_t,t}}\right] \leq d\left[c_1 m \log\left(1+\frac{\lambda_1(\Sigma_0)n}{\sigma^2 d}\right) + c_2 c \log\left(1+\frac{\lambda_1(\Sigma_q)m}{\lambda_d(\Sigma_0)}\right)\right],$$

which yields the desired result after we substitute this bound into Lemma 1. To simplify presentation in the main paper, c_1 and c_2 in Theorem 3 include the above logarithmic terms that multiply them.

C Proof of Theorem 4

From Assumption 1, there exists a basis of d actions such that if all actions in the basis are taken in task s by round t, it is guaranteed that $\lambda_d(G_{s,t}) \geq \eta/\sigma^2$. We modify `HierTS` to takes these actions first in any task s. Let $\mathcal{C}_t = \{s \in \mathcal{S}_t : \lambda_d(G_{s,t}) \geq \eta/\sigma^2\}$ be the set of *sufficiently-explored tasks* by round t.

Using $\mathcal{C}_t$, we decompose the Bayes regret as

$$\mathcal{BR}(m,n) \leq \mathbb{E}\left[\sum_{t\geq 1}\sum_{s\in\mathcal{S}_t} \mathbb{1}\{s \in \mathcal{C}_t\}\left(A_{s,*}^\top \theta_{s,*} - A_{s,t}^\top \theta_{s,*}\right)\right] + \mathbb{E}\left[\sum_{t\geq 1}\sum_{s\in\mathcal{S}_t} \mathbb{1}\{s \notin \mathcal{C}_t\}\left(A_{s,*}^\top \theta_{s,*} - A_{s,t}^\top \theta_{s,*}\right)\right].$$

For any task s and round t, we can trivially bound

$$\mathbb{E}\left[(A_{s,*} - A_{s,t})^\top \theta_{s,*}\right] \leq \mathbb{E}\left[\|A_{s,*} - A_{s,t}\|_{\hat{\Sigma}_{s,1}} \|\theta_{s,*}\|_{\hat{\Sigma}^{-1}_{s,1}}\right] \leq 2\sigma_{\max}\left(\|\mu_q\|_{\hat{\Sigma}^{-1}_{s,1}} + \mathbb{E}\left[\|\theta_{s,*} - \mu_q\|_{\hat{\Sigma}^{-1}_{s,1}}\right]\right),$$

where $\sigma_{\max} = \sqrt{\lambda_1(\Sigma_q + \Sigma_0)}$ as in Appendix B. Here we use that $\|A_{s,*} - A_{s,t}\|_2 \leq 2$ and that the prior covariance of $\theta_{s,*}$ is $\hat{\Sigma}_{s,1} = \Sigma_q + \Sigma_0$. We know from (2) that $\theta_{s,*} - \mu_q \sim \mathcal{N}(\mathbf{0}, \Sigma_q + \Sigma_0)$. This means that $\hat{\Sigma}_{s,1}^{-\frac{1}{2}}(\theta_{s,*} - \mu_q)$ is a vector of d independent standard normal variables. It follows that

$$\mathbb{E}\left[\|\theta_{s,*} - \mu_q\|_{\hat{\Sigma}^{-1}_{s,1}}\right] = \mathbb{E}\left[\|\hat{\Sigma}_{s,1}^{-\frac{1}{2}}(\theta_{s,*} - \mu_q)\|_2\right] \leq \sqrt{\mathbb{E}\left[\|\hat{\Sigma}_{s,1}^{-\frac{1}{2}}(\theta_{s,*} - \mu_q)\|_2^2\right]} = \sqrt{d}\,.$$

Since $s \notin \mathcal{C}_t$ occurs at most d times for any task s, the total regret due to forced exploration is bounded as

$$\mathbb{E}\left[\sum_{t\geq 1}\sum_{s\in\mathcal{S}_t} \mathbb{1}\{s \notin \mathcal{C}_t\}\left(A_{s,*}^\top \theta_{s,*} - A_{s,t}^\top \theta_{s,*}\right)\right] \leq 2\sigma_{\max}\left(\|\mu_q\|_{\hat{\Sigma}^{-1}_{s,1}} + \sqrt{d}\right) dm = c_3\,.$$

It remains to bound the first term in $\mathcal{BR}(m,n)$. On event $s \in \mathcal{C}_t$, `HierTS` samples from the posterior and behaves exactly as Algorithm 1. Therefore, we only need to bound $\mathcal{V}(m,n) = \mathbb{E}\left[\sum_{t\geq 1}\sum_{s\in\mathcal{S}_t} \mathbb{1}\{s \in \mathcal{C}_t\}\|A_{s,t}\|^2_{\hat{\Sigma}_{s,t}}\right]$ and then substitute the bound into Lemma 1. By the total covariance decomposition in Lemma 2, we have

$$\|A_{s,t}\|^2_{\hat{\Sigma}_{s,t}} = A_{s,t}^\top \tilde{\Sigma}_{s,t} A_{s,t} + A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_t\Sigma_0^{-1}M^{-1}A_{s,t}\,, \tag{11}$$

where $M = \Sigma_0^{-1} + G_{s,t}$ to reduce clutter. As in Appendix B, we bound the contribution of each term separately.

C.1 First Term in (11)

This term depends only on $\tilde{\Sigma}_{s,t}$, which does not depend on interactions with other tasks than task s. Therefore, the bound is the same as in the sequential case in Appendix B.1,

$$\sum_{t\geq 1}\sum_{s\in\mathcal{S}_t} \mathbb{1}\{s \in \mathcal{C}_t\} A_{s,t}^\top \tilde{\Sigma}_{s,t} A_{s,t} \leq c_1 dm \log\left(1+\frac{\lambda_1(\Sigma_0)n}{\sigma^2 d}\right),$$

where c_1 is defined in Appendix B.

C.2 Second Term in (11)

The difference from the sequential setting is in how we bound the second term in (11). Before we had $|\mathcal{S}_t| = 1$, while now we have $|\mathcal{S}_t| \le L \le m$ for some L. Since more than one task is acted upon per round, the telescoping identity in (10) no longer holds. To remedy this, we reduce the concurrent case to the sequential one. Specifically, suppose that task $s \in \mathcal{S}_t$ in round t has access to the concurrent observations of prior tasks in round t, for some order of tasks $\mathcal{S}_t = \{S_{t,i}\}_{i=1}^{L}$. As Theorem 3 holds for any order, we choose the order where sufficiently-explored tasks $s \in \mathcal{C}_t$ appear first.

Let $\mathcal{S}_{t,i} = \{S_{t,j}\}_{j=1}^{i-1}$ be the first $i-1$ tasks in $\mathcal{S}_t$ according to our chosen order. For $s = S_{t,i}$, let

$$\bar{\Sigma}_{s,t}^{-1} = \Sigma_q^{-1} + \sum_{z \in \mathcal{S}_{t,i}} (\Sigma_0 + G_{z,t+1}^{-1})^{-1} + \sum_{z \in [m] \setminus \mathcal{S}_{t,i}} (\Sigma_0 + G_{z,t}^{-1})^{-1}$$

be the reciprocal of the hyper-posterior covariance updated with concurrent observations in tasks $\mathcal{S}_{t,i}$. Next we show that $\bar{\Sigma}_t$ and $\bar{\Sigma}_{s,t}$ are similar.

Lemma 5. *Fix round t and $i \in [L]$. Let $s = S_{t,i}$ and $\lambda_d(G_{s,t}) \ge \eta/\sigma^2$. Then*

$$\lambda_1(\bar{\Sigma}_{s,t}^{-1}\bar{\Sigma}_t) \le 1 + \frac{\sigma^{-2}\lambda_1(\Sigma_q)(\lambda_1(\Sigma_0) + \sigma^2/\eta)}{\lambda_1(\Sigma_q) + (\lambda_1(\Sigma_0) + \sigma^2/\eta)/L}\,.$$

Proof. Using standard eigenvalue inequalities, we have

$$\lambda_1(\bar{\Sigma}_{s,t}^{-1}\bar{\Sigma}_t) = \lambda_1((\bar{\Sigma}_t^{-1} + \bar{\Sigma}_{s,t}^{-1} - \bar{\Sigma}_t^{-1})\bar{\Sigma}_t) \le 1 + \lambda_1((\bar{\Sigma}_{s,t}^{-1} - \bar{\Sigma}_t^{-1})\bar{\Sigma}_t) \le 1 + \frac{\lambda_1(\bar{\Sigma}_{s,t}^{-1} - \bar{\Sigma}_t^{-1})}{\lambda_d(\bar{\Sigma}_t^{-1})}\,. \tag{12}$$

By Weyl's inequalities, and from the definition of $\bar{\Sigma}_t$, we have

$$\begin{aligned}
\lambda_d(\bar{\Sigma}_t^{-1}) &\ge \lambda_d(\Sigma_q^{-1}) + \sum_{z \in [m]} \lambda_d((\Sigma_0 + G_{z,t}^{-1})^{-1}) = \lambda_d(\Sigma_q^{-1}) + \sum_{z \in [m]} \lambda_1^{-1}(\Sigma_0 + G_{z,t}^{-1}) \\
&\ge \lambda_d(\Sigma_q^{-1}) + \sum_{z \in [m]} (\lambda_1(\Sigma_0) + \lambda_1(G_{z,t}^{-1}))^{-1} \ge \lambda_d(\Sigma_q^{-1}) + (i-1)(\lambda_1(\Sigma_0) + \sigma^2/\eta)^{-1}\,.
\end{aligned}$$

In the last inequality, we use that the previous $i-1$ tasks $\mathcal{S}_{t,i}$ are sufficiently explored. Analogously to (10),

$$\begin{aligned}
\bar{\Sigma}_{s,t}^{-1} - \bar{\Sigma}_t^{-1} &= \sum_{z \in \mathcal{S}_{t,i}} (\Sigma_0 + (G_{z,t} + \sigma^{-2}A_{z,t}A_{z,t}^\top)^{-1})^{-1} - (\Sigma_0 + G_{z,t}^{-1})^{-1} \\
&= \sigma^{-2} \sum_{z \in \mathcal{S}_{t,i}} \Sigma_0^{-1} M_{z,t}^{-1} \frac{A_{z,t}A_{z,t}^\top}{1 + \sigma^{-2}A_{z,t}^\top M_{z,t}^{-1} A_{z,t}} M_{z,t}^{-1}\Sigma_0^{-1}\,,
\end{aligned}$$

where $M_{z,t} = \Sigma_0^{-1} + G_{z,t}$ to reduce clutter. Moreover, since $\|A_{z,t}\|_2 \le 1$ and $\sigma^{-2}A_{z,t}^\top M_{z,t}^{-1} A_{z,t} \ge 0$, we have

$$\lambda_1(\bar{\Sigma}_{s,t}^{-1} - \bar{\Sigma}_t^{-1}) \le (i-1)\sigma^{-2}\,.$$

Finally, we substitute our upper bounds to the right-hand side of (12) and get

$$\frac{\lambda_1(\bar{\Sigma}_{s,t}^{-1} - \bar{\Sigma}_t^{-1})}{\lambda_d(\bar{\Sigma}_t^{-1})} \le \frac{(i-1)\sigma^{-2}}{\lambda_1^{-1}(\Sigma_q) + (i-1)(\lambda_1(\Sigma_0) + \sigma^2/\eta)^{-1}} \le \frac{\sigma^{-2}\lambda_1(\Sigma_q)(\lambda_1(\Sigma_0) + \sigma^2/\eta)}{\lambda_1(\Sigma_q) + (\lambda_1(\Sigma_0) + \sigma^2/\eta)/L}\,,$$

where we use that the ratio is maximized when $i - 1 = L$. This completes the proof. □

Now we return to (11). First, we have that

$$\begin{aligned}
A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_t\Sigma_0^{-1}M^{-1}A_{s,t} &= A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_{s,t}^{\frac{1}{2}}\left(\bar{\Sigma}_{s,t}^{-\frac{1}{2}}\bar{\Sigma}_t^{\frac{1}{2}}\bar{\Sigma}_t^{\frac{1}{2}}\bar{\Sigma}_{s,t}^{-\frac{1}{2}}\right)\bar{\Sigma}_{s,t}^{\frac{1}{2}}\Sigma_0^{-1}M^{-1}A_{s,t} \\
&\le \lambda_1(\bar{\Sigma}_{s,t}^{-\frac{1}{2}}\bar{\Sigma}_t^{\frac{1}{2}}\bar{\Sigma}_t^{\frac{1}{2}}\bar{\Sigma}_{s,t}^{-\frac{1}{2}})A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_{s,t}\Sigma_0^{-1}M^{-1}A_{s,t} \\
&\le \lambda_1(\bar{\Sigma}_{s,t}^{-1}\bar{\Sigma}_t)A_{s,t}^\top M^{-1}\Sigma_0^{-1}\bar{\Sigma}_{s,t}\Sigma_0^{-1}M^{-1}A_{s,t}\,,
\end{aligned}$$

where we use that the above expression is a quadratic form. Next we apply Lemma 5 and get

$$\lambda_1(\bar{\Sigma}_{s,t}^{-1}\bar{\Sigma}_t) \leq 1 + \frac{\sigma^{-2}\lambda_1(\Sigma_q)(\lambda_1(\Sigma_0)+\sigma^2/\eta)}{\lambda_1(\Sigma_q)+(\lambda_1(\Sigma_0)+\sigma^2/\eta)/L} = c_4\,.$$

After $\bar{\Sigma}_t$ is turned into $\bar{\Sigma}_{s,t}$, we follow Appendix B.2 and get that the hyper-parameter regret is

$$c_2 c_4 c d \log\left(1 + \frac{\lambda_1(\Sigma_q)m}{\lambda_d(\Sigma_0)}\right),$$

where the only difference is the extra factor of c_4. Finally, we combine all upper bounds and get

$$\mathcal{V}(m,n) = \mathbb{E}\left[\sum_{t\geq 1}\sum_{s\in\mathcal{S}_t} \mathbb{1}\{s\in\mathcal{C}_t\}\,\|A_{s,t}\|^2_{\hat{\Sigma}_{s,t}}\right] \leq d\left[c_1 m\log\left(1+\frac{\lambda_1(\Sigma_0)n}{\sigma^2 d}\right) + c_2 c_4 c\log\left(1+\frac{\lambda_1(\Sigma_q)m}{\lambda_d(\Sigma_0)}\right)\right],$$

which yields the desired result after we substitute it into Lemma 1. To simplify presentation in the main paper, c_1 and c_2 in Theorem 4 include the above logarithmic terms that multiply them.

D Gaussian Bandit Regret Bounds

Our regret bounds in Section 6 can be specialized to K-armed Gaussian bandits (Section 4.1). Specifically, when the action set $\mathcal{A} = \{e_i\}_{i\in[K]}$ is the standard Euclidean basis in $\mathbb{R}^K$, Theorems 3 and 4 can be restated as follows.

Theorem 6 (Sequential Gaussian bandit regret)**.** *Let $|\mathcal{S}_t| = 1$ for all rounds t. Let $\delta = 1/(mn)$. Then the Bayes regret of* `HierTS` *is*

$$\mathcal{BR}(m,n) \leq \sqrt{2Kmn[c_1 m + c_2]\log(mn)} + c_3\,,$$

where $c_3 = O(K)$,

$$c_1 = \frac{\sigma_0^2}{\log(1+\sigma^{-2}\sigma_0^2)}\log\left(1+\frac{\sigma_0^2 n}{\sigma^2 K}\right), \quad c_2 = \frac{\sigma_q^2 c}{\log(1+\sigma^{-2}\sigma_q^2)}\log\left(1+\frac{\sigma_q^2 m}{\sigma_0^2}\right), \quad c = 1+\frac{\sigma_0^2}{\sigma^2}\,.$$

The main difference from the proof of Theorem 3 is that we start with the finite-action bound in Lemma 1. Other than that, we use the facts that $\lambda_1(\Sigma_0) = \lambda_d(\Sigma_0) = \sigma_0^2$ and $\lambda_1(\Sigma_q) = \sigma_q^2$.

Theorem 7 (Concurrent Gaussian bandit regret)**.** *Let $|\mathcal{S}_t| \leq L \leq m$. Let $\delta = 1/(mn)$. Then the Bayes regret of* `HierTS` *is*

$$\mathcal{BR}(m,n) \leq \sqrt{2Kmn[c_1 m + c_2]\log(mn)} + c_3\,,$$

where c_1 and c are defined as in Theorem 6,

$$c_2 = \frac{\sigma_q^2 c_4 c}{\log(1+\sigma^{-2}\sigma_q^2)}\log\left(1+\frac{\sigma_q^2 m}{\sigma_0^2}\right), \quad c_4 = 1+\frac{\sigma^{-2}\sigma_q^2(\sigma_0^2+\sigma^2)}{\sigma_q^2+(\sigma_0^2+\sigma^2)/L}\,,$$

and $c_3 = O(Km)$.

When we specialize Theorem 4, we note that $\eta = 1$, since the action set $\mathcal{A}$ is the standard Euclidean basis.

E Image Classification Experiment

We conduct an additional experiment that considers online classification using a real-world image dataset. The problem is cast as a multi-task linear bandit with Bernoulli rewards. Specifically, we construct a set of tasks where one image class is selected randomly to have high reward. In each task, at every round, K images are uniformly sampled at random as actions, and the aim of the learning agent is to select an image from the

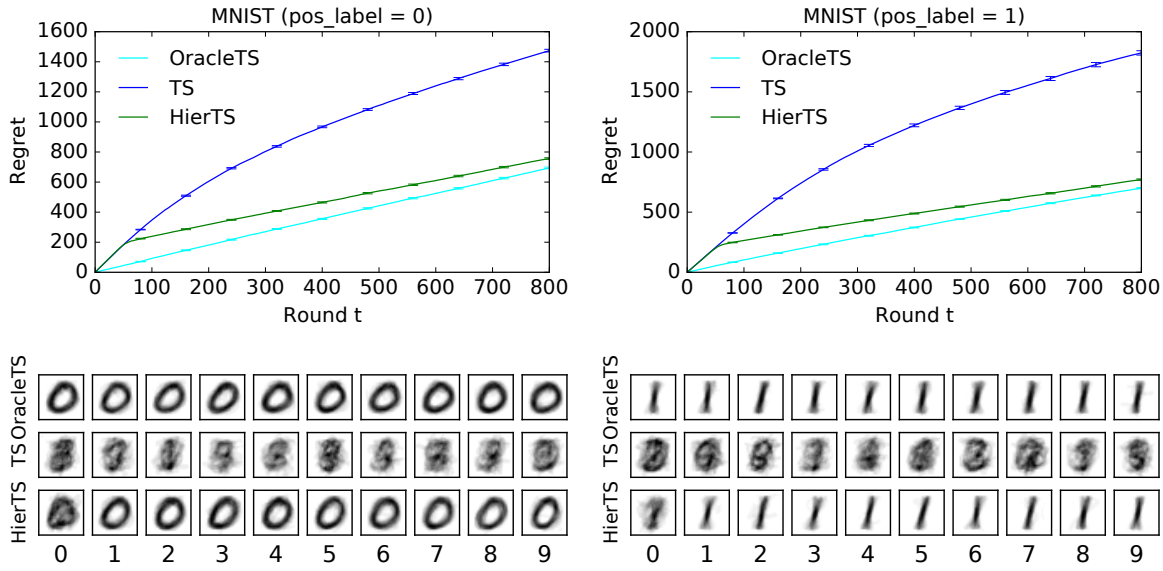

Figure 3: Evaluation of `HierTS` on multi-task digit classification using MNIST with different positive image classes. On the top, we plot the cumulative Bayes regret at each round. On the bottom, we visualize the most-rewarding image according to the learned hyper-parameter at evenly-spaced intervals.

unknown positive image class. The reward of an image from the positive class is $\mathsf{Ber}(0.9)$ and for all other classes is $\mathsf{Ber}(0.1)$.

We use the MNIST dataset (LeCun et al., 2010), which consists of $60,000$ images of handwritten digits, which we split equally into a training and test set. We down-sample each image to $d = 49$ pixels, which become the feature vector for the corresponding action in the bandit problem. For each digit, the training set is used to estimate μ_*, Σ_0, where all three algorithms use Σ_0 but only `OracleTS` can use μ_*. The algorithms are evaluated on the test set. Given a positive digit class, we construct a different task s by sub-sampling from the test set, and computing $\theta_{s,*}$ using positive images from the sub-sampled data. For each digit as the positive image class, we evaluate our three algorithms on a multi-task linear bandit with $m = 10$ tasks, $n = 400$ interactions per task, and $K = 30$ actions, uniformly sampled from the test images. We chose $L = 5$ tasks per round, leading to 800 rounds total. We assume a hyper-prior of $Q = \mathcal{N}(\mathbf{0}, I_d)$ with reward noise $\sigma = 0.5$ because the rewards are Bernoulli.

Figure 3 shows the performance of all algorithms for two digits across 20 independent runs. We see that `HierTS` performs very well compared to standard `TS`. In addition to regret, we also visualized the learned hyper-parameter $\bar{\mu}_t$ every 80 rounds. We see that `HierTS` very quickly learns the correct hyper-parameter, showing that it effectively leverages the shared structure across task. Overall, this experiment shows that even if `HierTS` assumes a misspecified model of the environment, with non-Gaussian rewards and not knowing the true hyper-prior Q and covariance Σ_0, `HierTS` still performs very well.