Hierarchical Thompson Sampling for Multi-band Radio Channel Selection

Jerrod Wigmore*, Brooke Shrader†, and Eytan Modiano*
*Massachusetts Institute of Technology, Cambridge, MA
†MIT Lincoln Laboratory, Lexington, MA

***Abstract*—We consider the multi-band channel selection problem, where the best channel is to be selected from n distinct frequency bands, each containing m wireless channels. The objective is to select the channel with the best average signal-to-interference-plus-noise ratio (SiNR), where the SiNR for each channel follows a parametric distribution, generated from a band-dependent prior distribution. We introduce a Bayesian Hierarchical Bandit (BHB) model that captures the correlation induced by the hierarchical relationship between channels and band, and develop a Hierarchical Thompson sampling (HTS) algorithm which leverages the underlying Bayesian Hierarchical structure to efficiently determine which channel is optimal. We demonstrate that the HTS algorithm outperforms traditional bandit algorithms by a factor of n when the bands are sufficiently dissimilar. Through extensive simulation, we characterize the Bayesian regret of the HTS algorithm under varying degrees of band similarity and demonstrate that the Bayesian regret of HTS does not increase linearly with n, in contrast to traditional bandit algorithms.**

I. Introduction

Wireless communication networks are increasingly moving beyond static spectrum allocations in favor of dynamic spectrum access over multiple frequency bands. For example, the 5G NR frequency bands include low-band spectrum below 1GHz, like 600 MHz and 700 MHz, mid-band spectrum between 1-6 GHz, such as 3.5 GHz and 4.9 GHz, and high-band spectrum above 6 GHz, such as 24 GHz, 28 GHz, and 39 GHz [1]. Multi-band radio communication is also utilized in both public safety and military networks, where the use of multiple radio access technologies, such as satellite communication in conjunction with terrestrial wireless communication, enhances the wireless network's flexibility and ability to withstand disruptions [2]. Dynamic Channel Selection (DCS) is a cognitive networking paradigm that enables radios to adapt to the wireless propagation and noise environment by dynamically selecting different channels in order to learn the optimal channel [3]. The goal of DCS is to improve the overall performance of the wireless communication system by maximizing the throughput, minimizing the interference, and increasing the reliability of communication [4].

In the single-band DCS problem, all channels share the same frequency band and are subject to the same propagation conditions. In the Multi-band DCS problem, channels are grouped into separate frequency bands as shown in Figure 1. Channels in the same band share the same propagation path and experience similar environmental conditions, however, there may still be variation in channel quality due to small-scale fading and interference. On the other hand, channels in different frequency bands may experience vastly different large-scale fading effects such as free-space path loss and shadowing, resulting in a bigger difference in channel quality between channels in different bands compared to channels within the same band. This correlation provides additional structure, and the information this structure provides can be exploited to improve the channel selection process. Additionally, prior knowledge regarding the quality of each band may provide valuable knowledge that may be leveraged in the multi-band DCS problem. This prior belief may be based on historical data or from simulations of the propagation environment. Due to these differences between the single-band and multi-band DCS problems, previous models and algorithms for the single-band case are insufficient for the multi-band DCS problem considered in this paper.

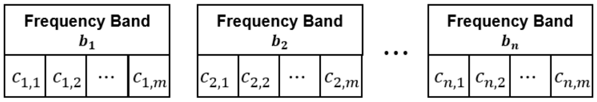


Fig. 1: Grouping of $n \times m$ individual channels $c_{i,j}$ into n distinct frequency bands

Multi-armed bandit (MAB) problems have been widely studied in the field of cognitive networking. In an MAB problem, a decision agent interacts with multiple independent arms or actions over multiple rounds. In each round, the agent chooses an arm and receives a random reward drawn i.i.d. from the arm's reward distribution. The objective is to maximize the expected cumulative reward over the problem's time horizon. The reward distributions are not known a priori, and thus the agent must balance *exploration*: learning about each arm reward distribution, and *exploitation*: selecting the arm(s) with the greatest estimated expected reward to maximize the expected cumulative reward. The MAB framework can be applied to DCS problems with ease: a cognitive radio selects channels sequentially, evaluating their quality using metrics

This work was funded in part by NSF grants CNS-2148183 and CNS-1907905. This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

ISBN 978-3-903176-57-7© 2023 IFIP

such as interference or the signal-to-interfere-plus-noise ratio (SiNR), with the aim of identifying the optimal channel. MAB frameworks have been used for other problems in the cognitive networking literature such as dynamic spectrum access: where spectrum resources are dynamically accessed by users of a shared spectrum [4]–[6], wireless scheduling: where channels are assigned to one or more users dynamically by a centralized entity or in a distributed manner [7], [8], and for self-organizing networks: where a network dynamically optimizes its topology and resources based on current traffic demands and network conditions [9], [10]. MAB problems have also been used to model sequential decision making in other domains including clinical trials, recommendation systems, and finance. For an overview on MAB problems, their variations, and applications, readers are directed to [11], [12].

The standard MAB model assumes arm reward distributions are statistically independent. However, in multi-band radio communication we cannot assume all channels are independent as channel quality will have band-dependent correlation. If the correlation between channels is known or can be learned, the correlation can be leveraged to improve the channel selection process. In this work we develop a Bayesian Hierarchical Bandit (BHB) model to capture correlation between channels in the multi-band DCS problem. The BHB model is an extension of Bayesian hierarchical modeling into the MAB problem domain. Bayesian hierarchical modeling uses Bayesian probability theory to model the relationships between different levels of a hierarchy, with each level representing a different level of abstraction or granularity. This approach allows for the estimation of unknown parameters by borrowing information across different levels of the hierarchy [13, chapter 5].

In the BHB model, *channels* correspond to arms of the MAB problem and are grouped into distinct frequency *bands*. We naturally assume each channel can only belong to a single band. The channel SiNR, or reward, distributions are parametric with a unique parameter for each channel. Each channel has an associated parametric *prior distribution* and the parameters for each channel within the band are sampled from its band's prior distribution. For this paper we focus on a specific model instance where the priors and SiNRs are Gaussian distributed with unknown means. The agent begins with knowledge of the parametric forms of the SiNR and prior distributions, but the agent does not have the knowledge of their respective parameters. During each time-step, the agent probes a single channel from one of the bands and measures the SiNR of that channel, which is obtained from that channel's underlying SiNR distribution. The SiNR it observes provides information not only on the chosen channel's parameter, but also the parameter of its associated band. As a result of the relationship between the band's prior distribution and its associated channels, the agent indirectly obtains information about all other channels within the band. While this is the first paper to consider the BHB model in the context of DCS, similar MAB extensions of Bayesian hierarchical models have been applied outside of the cognitive networking literature [14]–[18].

In addition to introducing this BHB model, we develop an algorithm that exploits this hierarchical structure to minimize the finite-time Bayesian regret, a metric which captures the expected regret of an algorithm with respect to the prior distribution. We show that when the band distributions are sufficiently far apart, this algorithm improves upon the Bayesian regret of classical MAB algorithms by a factor of n, where n is the number of bands. Our contributions are as follows: i.) We develop a novel Bayesian Hierarchical Bandit model that captures interdependence between channels seen in multi-band DCS problems. To the best of our knowledge, this is the first work that utilizes Bayesian hierarchical modeling in the cognitive radio/networking literature. ii.) We develop the Hierarchical Thompson sampling (HTS) algorithm which extends the posterior sampling framework of Thompson sampling (TS) to Bayesian hierarchical models and provide theoretical justification for its improved performance over TS. iii.) We demonstrate through simulation how the HTS algorithm outperforms TS in numerous settings, and characterize the problem settings where the performance gains are largest.

The remainder of this paper is outlined as follows. Section II provides our Bayesian Hierarchical Bandit model. In Section III, we present our extension of Thompson sampling, the Hierarchical Thompson sampling Algorithm, which leverages the hierarchical structure of the BHB model. Section IV demonstrates the empirical performance of these algorithms through simulation. Section V concludes the paper.

II. Model

We start by introducing notation. We use bold font to indicate vectors $\mathbf{b} = \{b_1, b_2, ..., b_k\} = \{b_i\}_{i\in[k]}$ where for positive integer k, $[k] = \{1, 2, ..., k\}$. The i-th element of vector $\mathbf{v}$ is represented as v_i. If a vector $\mathbf{v}_i$ is already indexed, the we denote its j-th element by $v_{i,j}$. We use upper-case symbols to indicate random variables or vectors e.g. Θ or $\mathbf{\Theta}$, and the corresponding lower-case symbols to indicate realizations of a random variables or random vectors e.g. θ or $\boldsymbol{\theta}$.

The Bayesian Hierarchical Bandit (BHB) model is as follows. There are n bands $\mathbf{b} = \{b_i\}_{i\in[n]}$, and each band contains m channels. The set of all channels belonging to band b_i is denoted as $\mathbf{c}_i = \{c_{i,j}\}_{j\in[m]}$, and the set of all channels across all bands is represented as $\mathbf{c} = \{\mathbf{c}_i\}_{i\in[n]}$. Each band b_i has its own band parameter θ_i and each channel $c_{i,j}$ has its own channel parameter $\theta_{i,j}$. The set of all channel parameters for $\mathbf{c}_i$ is denoted as $\boldsymbol{\theta}_i = \{\theta_{i,j}\}_{j\in[m]}$, and $\boldsymbol{\phi}_i = \{\theta_i, \boldsymbol{\theta}_i\}$ represents the set of all parameters associated with band b_i. The set of all parameters in the BHB model is represented as $\boldsymbol{\phi} = \{\boldsymbol{\phi}_i\}_{i\in[n]}$. For a problem instance, it is assumed that each parameter is produced through the following generative process: each band parameter $\{\theta_i\}_{i\in[n]}$ is generated independently from its band prior distribution $\pi_i = P(\theta_i)$. Each channel parameter $\theta_{i,j} \in \{\boldsymbol{\theta}_i\}_{i\in[n]}$ is generated independently from its channel prior distribution $\pi_{i,j} = P(\theta_{i,j}|\theta_i)$, which is parameterized

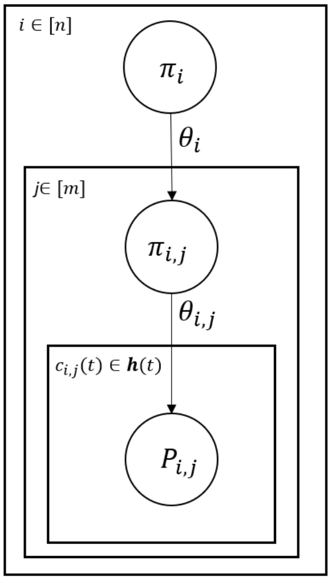


Fig. 2: Bayesian graphical model for the BHB model

by the corresponding band parameter θ_i. We use $P(\boldsymbol{\phi})$ to denote the joint prior distribution over all channel and band parameters in the BHB model.

In each of the $t \in [\tau]$ rounds, where τ is the time horizon, an agent probes a single channel $C(t) = c_{i,j}$ and measures an SiNR $s_{i,j}(t) \sim P_{i,j}$, where $P_{i,j} = P(s_{i,j}(t)|\theta_{i,j})$ is channel $c_{i,j}$'s SiNR distribution. Note that given $\theta_{i,j}$, the SiNR $S_{i,j}(t)$ is conditionally independent of band parameter θ_i. The band parameter θ_i only effects $S_{i,j}(t)$ through the channel parameter $\theta_{i,j}$. This Bayesian hierarchical model can be given concisely as:

$$\begin{aligned} \theta_i &\sim \pi_i \\ \theta_{i,j}|\theta_i &\sim \pi_{i,j}, && \forall\, j \in [m], \\ s_{i,j}(t)|\theta_{i,j} &\sim P_{i,j}, && \forall\, C(t) = c_{i,j} \end{aligned} \tag{1}$$

and is also visualized in Figure 2. All n band prior distributions are independent in our BHB model. This differs from the Bayesian hierarchical models used in the Meta-bandit literature in which all parameters are related by a common prior or root node in the Bayesian graphical model.

A. Parameter Estimation

Parameter estimation is crucial for the agent to exploit the hierarchical structure underlying the BHB model. The agent's *belief* on the parameters of the model $\boldsymbol{\phi}$ given the associated priors and observed history is captured by a *posterior distribution*. We denote the history by $\mathbf{h}(t) = \{(c(t'), s(t'))\}_{t'<t}$ which records all previous channel selections and observed SiNRs up to round t. We represent the history for channel $c_{i,j}$ and $\mathbf{c}_i$ as $\mathbf{h}_{i,j}(t) \in \mathbf{h}(t)$ and $\mathbf{h}_i(t) \in \mathbf{h}(t)$ respectively. We maintain independent posteriors for each $\boldsymbol{\phi}_i \in \boldsymbol{\phi}$ as all bands are independent from one another. The posterior distribution over all channel and band parameters for band b_i, represented as $P(\boldsymbol{\phi}_i|\mathbf{h}_i(t))$, is a multi-variate and potentially complex non-parametric distribution. However, the hierarchical relationship between the parameters allows us to take advantage of conditional independence, and express the posterior distributions as a product of simpler closed-form distributions:

$$\begin{aligned} P(\boldsymbol{\phi}_i|\mathbf{h}_i(t)) &= P(\theta_i, \boldsymbol{\theta}_i|\mathbf{h}_i(t)) \\ &= P(\theta_i|\mathbf{h}_i(t))P(\boldsymbol{\theta}_i|\mathbf{h}_i(t)) \\ &= P(\theta_i|\mathbf{h}_i(t)) \prod_{j\in[m]} P(\theta_{i,j}|\theta_i, \mathbf{h}_{i,j}(t)) \end{aligned} \tag{2}$$

When the band posterior $P(\theta_i|\mathbf{h}_i(t))$ and the channel posterior $P(\theta_{i,j}|\theta_i, \mathbf{h}_{i,j}(t)))$ are closed-form distributions, hierarchical sampling as described in Section III can be used to sample from the joint posterior distribution. Otherwise, inference and sampling requires approximation techniques [13].

B. Three-level Normal Model

For this paper we will focus on the Three-Level Normal (TLN) model. In the TLN model, the band prior, channel prior, and SiNR distributions are all normally distributed with unknown means and known variances. The prior model is described by:

$$\begin{aligned} \theta_i &\sim N(\kappa_i, \gamma_i^2) \\ \theta_{i,j}|\theta_i &\sim N(\theta_i, \lambda_i^2), && \forall\, j \in [m], \\ s_{i,j}(t)|\theta_{i,j} &\sim N(\theta_{i,j}, \sigma_s^2), && \forall\, C(t) = c_{i,j} \end{aligned} \tag{3}$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. For each band, the prior band variance γ_i^2 and prior channel variance λ_i^2 are assumed to be known by the agent. Additionally, the SiNR variance σ_s^2 is the same for all channels and is assumed to be known by the agent. This model exhibits hierarchical conjugancy as the band posteriors and channel posteriors are also normally distributed [13]. Under the TLN model, the band and channel posteriors exist as closed form distributions meaning they allow for exact sampling and are easily interpretable. Given history $\mathbf{h}_i(t)$, the band posterior distribution is:

$$P(\theta_i|\mathbf{h}_i(t)) = N(u_i(t), v_i^2(t)) \tag{4}$$

where the variance and mean parameters are

$$v_i^2(t) = \left(\frac{1}{\gamma_i^2} + \sum_{j\in[m]} \frac{1}{\lambda_i^2 + \frac{\sigma_s^2}{k_{i,j}(t)}} \right)^{-1} \tag{5}$$

$$u_i(t) = v_i^2(t) \left(\frac{\kappa_i}{\gamma_i^2} + \sum_{j\in[m]} \frac{\bar{s}_{i,j}(t)}{\lambda_i^2 + \sigma_s^2/k_{i,j}(t)} \right) \tag{6}$$

and

$$\bar{s}_{i,j}(t) = \frac{1}{k_{i,j}(t)} \sum_{s_{i,j}(t)\in\mathbf{h}_{i,j}(t)} s_{i,j}(t) \tag{7}$$

is the average of all previous SiNR measurements for channel c_{ij}. Here, $k_{i,j}(t) = |\mathbf{h}_{ij}(t)|$ is the number of times channel $c_{i,j}$ has been chosen up to time t. Equations (5) and (6) provide intuition on what algorithms would result in accurate estimations of the band parameter θ_i. The band posterior mean $u_i(t)$ is a weighted average of the prior mean κ_i and the average SiNR measurements $\bar{s}_{i,j}(t)$ for each channel within the band. The posterior mean is initially biased by the the prior mean parameter κ_i,and the weight of this bias is proportional to the prior variance γ_i^2. The weight of each $\bar{s}_{i,j}(t)$ term is proportional to the uncertainty. When $c_{i,j}$ has not been selected yet, $\bar{s}_{i,j}(t)$ does not contribute to the average. After being selected, the weight of $\bar{s}_{i,j}(t)$ increases from $1/(\lambda_i^2+\sigma_s^2)$ to $1/\lambda_i^2$ as $k_{i,j}(t) \to \infty$. Similarly, the posterior variance cannot be minimized by selecting only a single channel. Thus, to reduce the overall uncertainty in θ_i it is necessary to explore the various channels within the band. The necessary amount of exploration depends on the prior variance parameters γ_i^2 and λ_i^2.

The channel posterior distribution is:

$$P\left(\theta_{i,j}|\theta_i, \mathbf{h}_{i,j}(t)\right) = N\left(u_{i,j}(t), v_{i,j}^2(t)\right) \tag{8}$$

where

$$v_{i,j}^2(t) = \left(\frac{1}{\lambda_i^2} + \frac{k_{i,j}(t)}{\sigma_s^2}\right)^{-1} \tag{9}$$

$$u_{i,j}(t) = v_{i,j}^2(t)\left(\frac{\theta_i}{\lambda_i^2} + \frac{\bar{s}_{i,j}(t)}{\sigma_s^2/k_{i,j}(t)}\right) \tag{10}$$

Unlike the band posterior, the channel posterior variance only depends on the observations from the particular channel. Additionally, the channel posterior mean is a weighted average of the conditional band parameter θ_i and averaged observed SiNR $\bar{s}_{i,j}(t)$ for the channel. In practice, the band channel parameter θ_i is usually replaced by a sample $\hat{\theta}_i$ from the current band posterior. Like the band prior mean κ_i in the band posterior, θ_i acts a bias term and its influence on the channel posterior mean decays as confidence in the channel parameter grows. More specifically, as $k_{i,j}(t) \to \infty$, the channel posterior variance decays $v_{i,j}(t) \to 0$, and the channel posterior mean approaches the average observed SiNR for that channel $u_{i,j}(t) \to \bar{s}_{i,j}(t)$. The posterior parameters in equations (5), (6), (9) and (10) can be derived from [15, Appendix D].

C. Performance Measure

Denote a policy $\Psi : \mathbf{h}(t) \to c(t)$ as mapping from histories to channel probing decisions. For a given instance of parameters $\boldsymbol{\phi}$ and policy Ψ, the τ round *regret* is:

$$\mathcal{R}(\tau, \boldsymbol{\phi}, \Psi) = \mathbb{E}\left[\sum_{t=1}^{\tau} \mu^* - \mu(c(t))\middle|\boldsymbol{\phi}\right] \tag{11}$$

where $\mu(c(t))$ is the expected SiNR of channel chosen in round t, and $\mu^* = \max_{c_{i,j} \in \mathbf{c}} \mu(c_{i,j})$ is the max expected SiNR of any channel across all bands. The *Bayesian regret* for policy Ψ over τ rounds is defined as:

$$\mathcal{BR}(\tau, P, \Psi) = \mathbb{E}_{\boldsymbol{\phi} \sim P}\left[\mathcal{R}(\tau, \boldsymbol{\phi}, \Psi)\right] \tag{12}$$

where the expectation is taken with respect to the joint prior distribution $P(\boldsymbol{\phi})$. Bayesian regret is a weaker notion than *worst-case regret*:

$$\mathcal{R}(\tau, \Psi) = \max_{\boldsymbol{\phi}} \mathcal{R}(\tau, \boldsymbol{\phi}, \Psi) \tag{13}$$

in the sense that any bound on the worst-case regret implies a bound on the Bayesian regret as well [12]. However, in many real-world applications such as wireless channel selection, historical data or simulations may allow a decision maker to form a prior belief before interacting with the environment. In these applications, Bayesian regret is a good theoretical performance measure as it gives greater weight to likely problem instances compared to worst-case regret. Also, we focus on finite-time Bayesian regret as opposed to asymptotic regret as it best reflects the multi-band channel selection problem.

III. Algorithms

Thompson sampling (TS) is the oldest Multi-Armed bandit algorithm that dates back to 1933 [19]. However, TS was largely ignored until empirical studies demonstrated its efficiency [20] followed by theoretical performance guarantees within the past decade [21]. The core idea of Thompson sampling is to select an arm in each round according to the probability the chosen arm is optimal. Consider a non-hierarchical Bayesian bandit model with m independent arms, each parameterized by their individual arm parameter θ_j. Let $P(c^*|\mathbf{h}(t))$ be the *arm selection* distribution which is a discrete distribution over each arm. $P(c^* = c_j|\mathbf{h}(t))$ corresponds to the probability arm c_j has the greatest expected reward given history $\mathbf{h}(t)$. In each round, the selected arm $c(t)$ is sampled from $P(c^*|\mathbf{h}(t))$. This sampling step is typically performed by first sampling instances of arm parameters $\hat{\theta}_j(t)$ from the marginal arm posterior distributions $P(\theta_j|\mathbf{h}(t))$ for all $j \in [m]$. Then $c(t) = \operatorname{argmax}_{c_j \in \mathbf{c}} \mathbb{E}\left[\mu(c_j)|\hat{\theta}_j(t)\right]$ is the corresponding sample from $P(c^*|\mathbf{h}(t))$. The full TS algorithm for the non-hierarchical Bayesian bandit is given in Algorithm 1.

Algorithm 1 Thompson sampling

```
Require: $\pi_j = P(\theta_j)$ and $P(s_j(t) \mid \theta_j)\,\forall\, j \in [m]$
  for t=1,...,$\tau$ do
    for $j = 1, ..., m$ do
      Sample $\hat{\theta}_j \sim P(\theta_j|\mathbf{h}_j(t))$
    end for
    $c(t) = \operatorname{argmax} \mathbb{E}\left[\mu(c_j)|\hat{\theta}_j\right]$
    Select $c(t) = c_j$ and observe $s_j(t) \sim P(s_j(t)|\theta_j)$
    Update $P(\theta_j|\mathbf{h}(t))$
  end for
```

A. Hierarchical Thompson sampling

In this section we provide an extension of TS to the BHB model. Like in TS, we need the ability to sample from the channel selection distribution $P(c^*|\mathbf{h}(t))$. Unlike TS, in the BHB model the marginal channel posterior distribution is:

$$P(\theta_{i,j}|\mathbf{h}_i(t)) = \int_{\theta_i} P(\theta_{i,j}, \theta_i|\mathbf{h}_i(t))dP(\theta_i) \tag{14}$$

which is not explicitly defined in the model. Even in BHB models with closed-form band posteriors and channel posteriors, the marginal channel posterior distribution may be difficult to compute or may not exist in closed form. Thus Hierarchical Thompson sampling (HTS) uses sequential posterior sampling to generate $\hat{\theta}_{i,j}(t)$ as follows:

1) Sample $\hat{\theta}_i(t) \sim P(\theta_i|\mathbf{h}_i(t))$ for all $i \in [n]$
2) Sample $\hat{\theta}_{i,j}(t) \sim P(\theta_{i,j}|\hat{\theta}_i(t), \mathbf{h}_{i,j}(t))$ for all $i \in [n]$ and $j \in [m]$

This process is equivalent to drawing

$$\hat{\theta}_i(t), \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}(t) \sim P(\theta_i, \boldsymbol{\theta}_i|\mathbf{h}_i(t))$$

and thus the samples $\hat{\theta}_{i,j}(t) \in \hat{\boldsymbol{\theta}}_i$ are equivalently sampled according to their respective marginal posteriors $P(\theta_{i,j}|\mathbf{h}_i(t))$. As in TS, $c(t) = \operatorname{argmax}_{c_{i,j} \in \mathbf{c}} \mathbb{E}\big[\mu(c_{i,j})|\hat{\theta}_{i,j}(t)\big]$ is the corresponding sample from $P(c^*|\mathbf{h}(t))$.

Algorithm 2 Hierarchical Thompson sampling

```
Require: $\pi_i, \pi_{i,j}$ and $P(s_{i,j}(t) \mid \theta_{i,j}) \forall i, j$
  for t=1,..., $\tau$ do
    for $i = 1, ..., n$ do
      Sample $\hat{\theta}_i(t) \sim P(\theta_i|\mathbf{h}_i(t))$
      for $j = 1, ..., m$ do
        Sample $\hat{\theta}_{i,j}(t) \sim P(\cdot|\hat{\theta}_i, \mathbf{h}_{i,j}(t))$
      end for
    end for
    $c(t) = \text{argmax}\, \mathbb{E}[\mu(c_{i,j})|\hat{\theta}_{i,j}(t)]$
    Select $c(t) = c_{i,j}$ and observe $s_{i,j}(t) \sim P(s_{i,j}(t)|\theta_{i,j})$
    Update $P(\theta_i|\mathbf{h}_i(t))$ and $\mathbb{P}(\theta_{i,j}|\theta_i, \mathbf{h}_{i,j}(t))$
  end for
```

The HTS algorithm learns both the channel and band parameters. The knowledge of band parameters is then used for more efficient exploration of the channels. For example, say the agent quickly learned all $\theta_i \in \boldsymbol{\theta}$ for an instance of the TLN model. This means sufficiently many channels within each band have been chosen in previous rounds, and the band posterior variance $v_i^2(t)$ is small for each band. In subsequent rounds, the posterior samples $\hat{\theta}_i(t) \approx \theta_i$. As seen from Equation (10), the channel posterior distributions, and their samples, will be biased towards their respective band parameter $\theta_i(t)$ when $k_{i,j}(t)$ is low. As a result, once the band parameters are sufficiently learned, exploration will primarily be focused on channels within the best bands.

B. Performance Improvement

To analyze the performance of HTS, we compare it to the standard TS algorithm which does not take the hierarchical modeling into account. The comparison highlights the value of incorporating grouping information, e.g. a channel's membership to a band, in sequential decision making applications. For a given BHB model, we define the corresponding *TS bandit model* as one that considers all channels to be independent but incorporates the same marginal channel prior knowledge as the BHB model. The prior for each channel parameter in the TS bandit model is derived by marginalizing the joint prior of the BHB model over the band parameter.

$$P(\theta_{i,j}) = \int_{\theta_i} P(\theta_i)P(\theta_{i,j}|\theta_i)dP(\theta_i) \tag{15}$$
$$= N(\kappa_i, \gamma_i^2 + \lambda_i^2) \tag{16}$$

After marginalizing, all band specific dependencies are ignored between channels and all channel parameters are treated as independent in the TS bandit model. The corresponding TS bandit model for the TLN model is:

$$\begin{aligned} \theta_{i,j} &\sim N(\kappa_i, \gamma_i^2 + \lambda_i^2) \qquad && \forall\, i \in [n], j \in [m] \\ s_{i,j}(t)|\theta_{i,j} &\sim N(\theta_{i,j}, \sigma_s^2) && \forall\, C(t) = c_{i,j} \end{aligned} \tag{17}$$

The channel posteriors under the equivalent non-hierarchical model are independent and follow from the Normal conjugate prior model for unknown means [13, pg. 39]:

$$P(\theta_{i,j}|\mathbf{h}_{i,j}(t)) = N(\tilde{u}_{i,j}(t), \tilde{v}_{i,j}^2(t)) \tag{18}$$

where

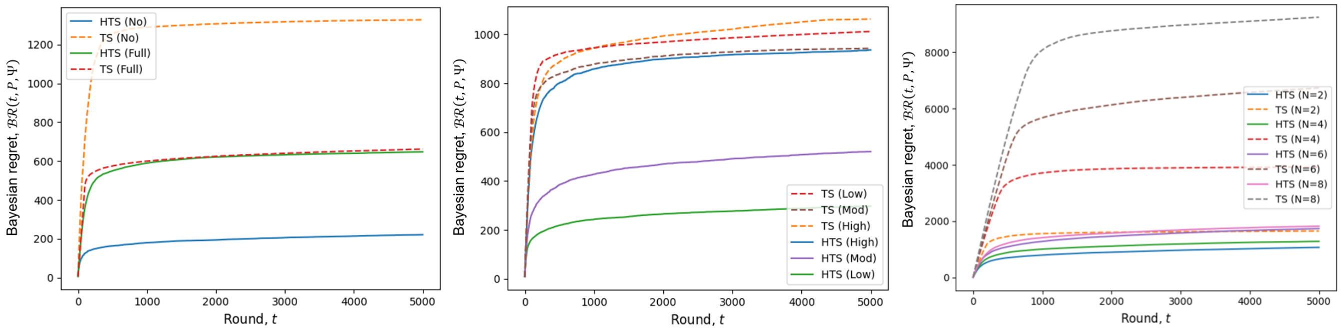
$$\tilde{v}_{i,j}^2 = \left(\frac{1}{\gamma_i^2 + \lambda_i^2} + \frac{k_{i,j}(t)}{\sigma_s^2}\right)^{-1} \tag{19}$$

$$\tilde{u}_{i,j}(t) = \tilde{v}_{i,j}^2(t)\left(\frac{\kappa_i}{\gamma_i^2 + \lambda_{i,j}^2} + \frac{\bar{s}_{i,j}(t)}{\sigma_s^2/k_{i,j}(t)}\right) \tag{20}$$

HTS improves upon TS in the case there is additional information provided by grouping of channels into bands. Its intuitively obvious that the greatest improvement of HTS compared to TS would occur in environments where the generative channel distributions $P(\theta_{i,j}|\Theta_i = \theta_i)$ for each band are non-overlapping. Additionally, when all band distributions are identical, we would expect there to be no advantage in using HTS. We formalize these intuitions in the the following claims.

Claim 1. *There exists TLN priors such that the Bayesian regret of HTS is the same as the Bayesian regret of TS*

Proof. Assume the band and channel parameters are generated from the the TLN model and let $\gamma_i = 0$, $\kappa_i = \kappa$, and $\lambda_i = \lambda$ for all $i \in [n]$. For environments generated from this prior model, all band parameters $\theta_i = \kappa$ are equivalent. Thus all channel parameters $\{\theta_{i,j}\}_{i \in [n], j \in [m]}$ are distributed according to $P(\theta_{i,j}) = N(\kappa, \lambda^2)$ which is equivalently the marginal prior that would be used by the the TS agent under $\gamma_i = 0$. Additionally, given $\mathbf{h}(t)$, the expected marginal posterior parameters $(\dot{u}_{i,j}(t), \dot{v}_{i,j}^2(t))$ for the HTS model are equivalent to the posterior parameters $(\tilde{u}_{i,j}(t), \tilde{v}_{i,j}^2(t))$ for TS model. Meaning the channel selection distribution $P(c^*|\mathbf{h}(t))$



(a) Average regret over 30 random environments from the "No" and "Full" overlap scenarios

(b) Average regret over 30 random environments from the low, moderate (mod), and high similarity prior settings.

(c) Average regret per round in four scenarios with a different number of bands but the same prior specification

Fig. 3: Average regret over 5000 rounds for all scenarios discussed in Section IV

.

is the same for the TS and HTS algorithm. It follows that the Bayesian regret is the same for both algorithms.

□

Claim 2. *There exists TLN priors such that the Bayesian regret of HTS is $O(n)$ improvement over the Bayesian regret of TS*

Proof. Assume the band and channel parameters are generated from the TLN model, and let $\kappa_i = \kappa$, $\gamma_i^2 = \gamma$, $\lambda_i^2 = \lambda^2$ for all $i \in [m]$. Furthermore, assume $\gamma^2 \gg \lambda^2$. If the difference between the band prior variance γ^2 and channel prior variance λ^2 is sufficiently large, after probing one channel from each band, all band posteriors $P(\theta_i|\mathbf{h}_i(t))$ will be non-overlapping. Under the HTS algorithm, after $O(n)$ rounds a single best band will be identified and further exploration will be constrained to channels within the best band. Under the same prior specification, all non-selected channels under the TS algorithm will have the same prior until being selected and exploration will not be constrained to a single band. Since TS is guaranteed to explore all nm channels, while the HTS algorithm will only explore $O(m)$ channels, the Bayesian regret of the HTS algorithm is an $O(n)$ improvement over the Bayesian regret of the TS algorithm.

□

Characterizing the Bayesian regret of HTS in terms of the number of channels/bands and the prior parameters is an open problem. The following upper-bound can be derived using the same methodology as in [21]:

$$\mathcal{BR}(\tau,P,\Psi) \le \mathbb{E}\left[\sqrt{\tau}\varepsilon(\tau)\sum_{t=1}^{\tau}\dot{v}_{i,j}^2(t)\right] + \max_{i\in[n]}(\gamma_i^2+\lambda_i^2) \quad (21)$$

where $\dot{v}_{i,j}^2(t)$ is the marginal posterior variance of the arm selected by the agent in round t and $\varepsilon(\tau) = \ln((\tau^2+1)nm/(\sqrt{2\pi}))$. An upper-bound on the Bayesian regret with respect the number of channels/bands and the prior parameters is not apparent from Equation (21). We conjecture that the Bayesian regret of HTS is of order $O(c(P(\boldsymbol{\phi}))m)$, where $c(P(\boldsymbol{\phi})) \in [1, n]$ is a function of the similarity between the band distribution for the band containing the best channel and the remaining cluster distributions based on the prior $P(\boldsymbol{\phi})$.

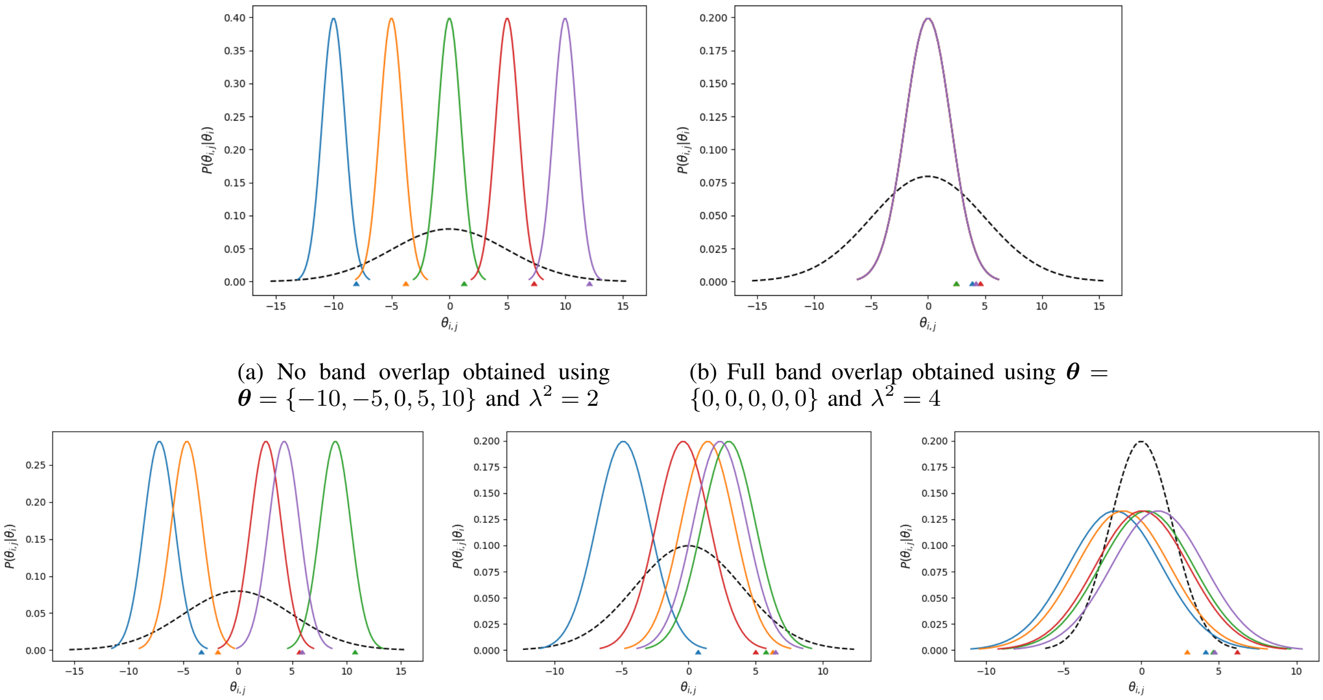
In the following section, we validate these claims through simulation and demonstrate empirically how the improvement in Bayesian regret is dependent on the average band similarity which is dictated by the prior distributions.

IV. Numerical Experiments

In this section we evaluate the empirical performance of HTS using simulations. First, we provide background on the generation process of problem instances used in the simulations. Then we empirically validate Claims 1 and 2 via non-overlapping and overlapping BHB problem instances respectively. Next, we compare the performance of HTS to TS in more general problem instances. Finally, we demonstrate how the performance improvement of HTS compared to TS scales with the number of bands n.

A. Simulation Setup

For all experiments, a TLN prior $P(\boldsymbol{\phi})$ is specified by setting the values of κ_i, γ_i^2 and λ_i^2 for all $i \in [n]$. Depending on the experiment, all band parameters $\{\theta_i\}_{i\in[n]}$ were either hand-picked or sampled from the environment's band prior distribution $P(\theta_i) = N(\kappa_i, \gamma_i^2)$. For all experiments, the channel parameters $\{\theta_{i,j}\}_{i\in[n],j\in[m]}$ were sampled from their respective channel parameter generating distributions $P(\theta_{i,j}|\Theta_i = \theta_i) = N(\theta_i, \lambda_i^2)$. We denote a problem instance as $\mathcal{I} = (P(\boldsymbol{\phi}), \boldsymbol{\phi})$. Both TS (Algorithm 1) and HTS (Algorithm 2) are run on each problem instance. The TS algorithm considers all channels to be independent and uses the corresponding TS bandit model as specified by equations (17)–(20). The HTS algorithm uses the TLN model as described by equations (3)–(6) and (8)–(10). For each experiment, we use the same prior parameters for all trials meaning $\kappa_i = \kappa$, $\gamma_i^2 = \gamma$, and $\lambda_i^2 = \lambda^2$ for all $i \in [n]$. Both HTS and TS algorithms are initialized with these parameters. Note, using the same prior parameters for each band does not mean all



(a) No band overlap obtained using $\boldsymbol{\theta} = \{-10, -5, 0, 5, 10\}$ and $\lambda^2 = 2$

(b) Full band overlap obtained using $\boldsymbol{\theta} = \{0, 0, 0, 0, 0\}$ and $\lambda^2 = 4$

(c) Low band overlap obtained using $\kappa = 0$, $\gamma^2 = 25$, and $\lambda^2 = 2$

(d) Moderate band overlap obtained using $\kappa = 0$, $\gamma^2 = 16$, and $\lambda^2 = 4$

(e) High band overlap obtained using $\kappa = 0$, $\gamma^2 = 2$, and $\lambda^2 = 9$

Fig. 4: Examples of channel generating distributions for the scenarios in Table I and Table II. The black dashed line in each subfigure is the band prior π_i for each cluster, and the colored lines represent the channel parameter generating distributions $P(\theta_{i,j}|\Theta_i = \theta_i)$ for each band. The colored arrows above the x-axes indicate the greatest expected channel SiNR $\max_{c_{ij} \in \mathbf{c}_i} \mu(c_{i,j})$ in each corresponding cluster.

generated band and channel parameters are identical, only that the algorithms start with the same prior distribution over the parameters.

1) Quantifying Overlap: To quantify the overlap between two channel parameter generating distributions, we use the overlap-coefficient (OVL) [22]:

$$OVL(f_i(x), f_{i'}(x) = \int_x \min(f_i(x), f_{i'}(x))dx \tag{22}$$

where $f_i(x)$ is the probability density function of $P(\theta_{i,j}|\Theta_i = \theta_i)$. The OVL is the fraction of the shared probability mass for the two densities and is always bounded by $[0, 1]$. The OVL measures the overlap between a single pair of distributions. As a measure of the amount of the average overlap in a problem instance $\mathcal{I} = (P(\boldsymbol{\phi}), \boldsymbol{\phi})$ with n bands, we define the *average overlap score* as:

$$V(\mathcal{I}) = \frac{1}{\mathcal{P}(n)} \sum_{\substack{i \in [n], \\ (i' > i) \in [n]}} OVL(f_i(x), f_{i'}(x)$$

where $\mathcal{P}(n) = \frac{n(n-1)}{2}$ is the number of unique pairwise OVL scores computed in the sum, and is included to bound $V(\mathcal{I}) \in [0, 1]$.

B. Numerical Validation of Claims

We begin by empirically validating Claims 1 and 2. For validating these claims we generated 30 TLN instances with $V(\mathcal{I}) = 0$ and 30 TLN instances with $V(\mathcal{I}) = 1$ respectively. To ensure the each instance generated had the intended average overlap score, we fixed the band parameters $\boldsymbol{\theta} = \{-10, -5, 0, 5, 10\}$ for all Claim 1 trials and $\boldsymbol{\theta} = \{0, 0, 0, 0, 0\}$ for all Claim 2 trial. Each band contained $m = 100$ channels, and the channel parameter generating distributions $P(\theta_{i,j}|\Theta_i = \theta_i)$ for these tests are shown in Figures 4a and 4b. For each TLN instance, HTS and TS were run for a total of $\tau = 5000$ rounds. The regret per round averaged all 30 problem instances for each experiment is plotted in Figure 3a, and the final average regret is found in Table I along with prior parameters used for each trial. Its clear that in the case there is no overlap, HTS outperforms TS by a factor greater than $n = 5$. When the channel parameter generating distributions are fully overlapping, there is no difference in the performance as HTS and TS are functionally equivalent.

C. Performance vs Overlap

Next we demonstrate the performance when the overlap lies between the two extremes. In the following experiments, the band parameters $\{\theta_i\}_{i \in [n]}$ were sampled from their band

Overlap Scenario	Avg. Cluster Overlap	κ	γ^2	λ^2	Average Regret		Factor Improvement
					HTS	TS	
No	0	0	25	2	220.80	1327.92	6.0
Full	1	0	25	4	647.49	661.94	1.0

TABLE I: Prior parameters and results for the No and Full overlap scenarios

priors $\{P(\theta_i)\}_{i\in[n]}$, meaning the generated band parameters were random and no longer identical for each problem instance within a single experiment. The amount of overlap between bands was controlled via the specification of the priors. We utilized three different sets of priors and reference the corresponding experiments as "Low", "Moderate", and "High" overlap scenarios. Each band contained $m = 100$ channels, and the prior parameters used to generate the TLN instances for each scenario are given in Table II. Once again, we generated 30 problem instances for each experiment and both algorithms were run for a total of $\tau = 5000$ rounds on each problem instance. Figure 3b plots the average regret per round for these experiments and the final average regret values are found in Table II. These results suggests that as the overlap between bands decreases, the performance improvement of HTS increases.

Overlap Scenario	Avg. Cluster Overlap	κ	γ^2	λ^2	Average Regret		Factor Improvement
					HTS	TS	
Low	0.32	0	25	2	297.44	1012.14	3.4
Moderate	0.61	0	16	4	543.96	908.29	1.7
High	0.89	0	4	9	936.59	1063.10	1.1

TABLE II: Prior parameters and results for the Low, Moderate, and High overlap scenarios

D. Number of bands comparison

Next we examine how the number of bands n effects the the magnitude of improvement of HTS over TS. For this comparison, we ran experiments with $n = 2$, 4, 6, and 8 bands per problem instance. For each experiment, the moderate overlap prior parameters were used ($\kappa_i = 0$, $\gamma_i^2 = 16$, and $\lambda_i^2 = 4$ for all $i \in [n]$), and each band contained $m = 100$ channels. Once again 30 different problem instances were generated for each experiment, and the results were averaged over these problem instances. Both HTS and TS were ran on each problem instance for a total of $\tau = 5000$ rounds. Figure 3c plots the average regret per round over all problem instances for each experiment. Table III provides the final average regret for each experiment. It is evident that the average regret of TS scales linearly with n while average regret of HTS only scales sub-linearly with n. As expected, this implies the largest gains in performance over TS would be seen in scenarios that have many well-separated bands.

V. Concluding Remarks

In this work, we introduced our Bayesian Hierarchical Bandit model for the DCS problem. We then provided a bandit algorithm, Hierarchical Thompson sampling, that leverages the known hierarchical structure to efficiently select channels to minimize the Bayesian regret. We provided theoretical justification for the improved performance of HTS over TS. Finally, we demonstrated empirically that HTS outperforms TS in numerous settings, with the performance gain being environment dependent. One possible direction of future work is to develop prior-dependent Bayesian regret bounds. These regret bounds would allow us to better understand the performance improvements over TS demonstrated empirically in Section IV. Extending the HTS algorithm to BHB models without closed-form posteriors is another promising future direction.

n	**Average Regret**		**Factor Improvement**
	HTS	**TS**	
2	1069	1650	1.5
4	1283	3922	3.1
6	1737	6733	3.9
8	1821	9244	5.1

TABLE III: $\tau = 5000$ round average regret for the scenarios plotted in Figure 3c

References

[1] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR: the Next Generation Wireless Access Technology*, 2020.

[2] B. Wang and K. J. R. Liu, "Advances in Cognitive Radio Networks: A Survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 5–23, 2011.

[3] K.-L. A. Yau, P. Komisarczuk, and P. D. Teal, "A Context-Aware and Intelligent Dynamic Channel Selection Scheme for Cognitive Radio Networks," *2009 4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications*, pp. 1–6, 2009.

[4] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive Medium Access: Exploration, Exploitation, and Competition," *IEEE Transactions on Mobile Computing*, vol. 10, no. 2, pp. 239–253, 2010.

[5] J. Zhu, Y. Song, D. Jiang, and H. Song, "Multi-Armed Bandit Channel Access Scheme With Cognitive Radio Technology in Wireless Sensor Networks for the Internet of Things," *IEEE Access*, vol. 4, pp. 4609–4617, 2016.

[6] N. Modi, P. Mary, and C. Moy, "QoS Driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-Armed Bandit Approach," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 1, pp. 49–66, 2017.

[7] F. Li, D. Yu, H. Yang, J. Yu, H. Karl, and X. Cheng, "Multi-Armed-Bandit-Based Spectrum Scheduling Algorithms in Wireless Networks: A Survey," *IEEE Wireless Communications*, vol. 27, no. 1, pp. 24–30, 2020.

[8] A. Alipour-Fanid, M. Dabaghchian, R. Arora, and K. Zeng, "Multiuser Scheduling in Centralized Cognitive Radio Networks: A Multi-Armed Bandit Approach," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1074–1091, 2022.

[9] T. Daher, S. B. Jemaa, and L. Decreusefond, "Cognitive Management of Self — Organized Radio Networks Based on Multi Armed Bandit," in *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, ser. 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017, pp. 1–5.

[10] M.-J. Youssef, V. V. Veeravalli, J. Farah, C. A. Nour, and C. Douillard, "Resource Allocation in NOMA-Based Self-Organizing Networks Using Stochastic Multi-Armed Bandits," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 6003–6017, 2021.

[11] T. Lattimore and C. Szepesvári, *Bandit Algorithms*, 1st ed. Cambridge University Press, 2020. [Online]. Available: https://www.cambridge.org/core/product/identifier/9781108571401/type/book

[12] A. Slivkins, "Introduction to Multi-Armed Bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[13] A. Gelman, J. Carlin, H. Stern, D. Rubin, D. Dunson, Vehtari, and Aki, *Bayesian Data Analysis*. CRC Press, 2013.

[14] R. Wan, L. Ge, and R. Song, “Metadata-based Multi-Task Bandits with Bayesian Hierarchical Models,” in *Advances in Neural Information Processing Systems*, ser. arXiv, vol. 34, 2021, pp. 29 655—29 668.

[15] B. Kveton, M. Konobeev, M. Zaheer, C.-w. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari, “Meta-Thompson Sampling,” in *International Conference on Machine Learning*, ser. PMLR, 2021, pp. 5884—5893.

[16] J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh, “Hierarchical Bayesian Bandits,” in *International Conference on Artificial Intelligence and Statistics*, ser. arXiv, 2022, pp. 7724—7741.

[17] J. Hong, B. Kveton, S. Katariya, M. Zaheer, and M. Ghavamzadeh, “Deep Hierarchy in Bandits,” in *International Conference on Machine Learning*, ser. arXiv. PMLR, 2022, pp. 8833—8851.

[18] R. Wan, L. Ge, and R. Song, “Towards Scalable and Robust Structured Bandits: A Meta-Learning Framework,” *arXiv*, 2022.

[19] W. R. Thompson, “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples,” *Biometrika*, vol. 25, no. 3/4, p. 285, 1933.

[20] O. Chapelle and L. Li, “An Empirical Evaluation of Thompson Sampling,” in *Advances in Neural Information Processing Systems*, vol. 24. Curran Associates, Inc., 2011.

[21] D. Russo and B. V. Roy, “Learning to Optimize via Posterior Sampling,” *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[22] H. F. Inman and E. L. Bradley, “The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities,” *Communications in Statistics - Theory and Methods*, vol. 18, no. 10, pp. 3851–3874, 1989.