Towards Scalable and Robust Structured Bandits: A Meta-Learning Framework

Runzhe Wan[*] **Lin Ge**[*] **Rui Song**
North Carolina State University

Abstract

Online learning in large-scale structured bandits is known to be challenging due to the curse of dimensionality. In this paper, we propose a unified meta-learning framework for a wide class of structured bandit problems where the parameter space can be factorized to item-level, which covers many popular tasks. Compared with existing approaches, the proposed solution is both scalable to large systems and robust by utilizing a more flexible model. At the core of this framework is a Bayesian hierarchical model that allows information sharing among items via their features, upon which we design a meta Thompson sampling algorithm. Three representative examples are discussed thoroughly. Theoretical analysis and extensive numerical results both support the usefulness of the proposed method.

1 INTRODUCTION

The bandit problem has received increasing attention and has been widely applied [Lattimore and Szepesvári, 2020]. However, many real-world applications typically have a large number of unknown parameters, a huge action space, and a complex reward distribution specified by domain models. For instance, in online learning to rank, the agent typically needs to choose a slate from more than thousands of related items [Li et al., 2016, Zong et al., 2016], and online advertising on major websites is usually viewed as a bipartite matching problem with millions of users and items [Wen et al., 2015]. *How to efficiently and reliably explore and learn in a large-scale structured bandit problem* is known to be challenging [Wen et al., 2015, Zong et al., 2016, Oh and Iyengar, 2019], which impedes the deployment of bandits in many real systems.

In this paper, we focus on a class of structured bandit problems where the parameter space can be factorized, with each parameter related to one item. Here, an item can be a product, a web page, a movie, etc., depending on the application. Every item typically also has an informative feature vector. Such a class is very general and includes many popular bandit problems as special cases, such as dynamic assortment optimization [Agrawal et al., 2017, 2019], online learning to rank [Kveton et al., 2015], online combinatorial optimization [Chen et al., 2013], multi-product dynamic pricing [Bastani et al., 2019], rank-1 bandits [Katariya et al., 2017], online revenue management [Ferreira et al., 2018], etc.

There are two major approaches dominating this area in the past decade [Chen et al., 2013, Kveton et al., 2015, Wen et al., 2015, Sankararaman, 2016, Li et al., 2016, Zong et al., 2016, Agrawal et al., 2017, Ferreira et al., 2018, Wang and Chen, 2018, Ou et al., 2018, Agrawal et al., 2019, Cheung et al., 2019, Bastani et al., 2019, Dong et al., 2020, Agrawal et al., 2020, Chen et al., 2021, Kveton et al., 2022], while both of them have limitations (see Section 3 for more details): the *feature-agnostic* approach learns every item from scratch and is therefore statistically *non-scalable*; while the *feature-determined* approach assumes we can use features to predict item-specific parameters perfectly with no error, and hence it relies on a fairly restricted (*non-robust*) model assumption to share information.

Intuitively, appropriate information sharing between items can largely speed up our learning, while a restricted generalization function may cause a linear regret due to the bias. To address these limitations, we propose a meta-learning framework: we first build a Bayesian hierarchical model to allow information sharing among items via their features, upon which we then design a Thompson sampling (TS, Russo et al. [2017])-type algorithm. The hierarchical model provides a principled way to construct a feature-based informative prior for each item, which guides the exploration of TS. As such, our method can be viewed as *learning how to learn* efficiently (i.e., *meta-*

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

[*] Equal contribution.

learning) for each item and hence for the whole problem, which improves the scalability. Compared with the feature-determined approach, ours allows the item-specific parameter to be only partially explained by its features, and hence is expected to be more robust with a more flexible model.

Contribution. Our contributions are multi-fold. First, to address the long-standing challenges in large-scale structured bandits, we propose a unified meta-learning framework with a TS-type algorithm, named Meta Thompson Sampling for Structured bandits (`MTSS`). To our knowledge, this is the *first* meta-learning approach to solve the wide class of structured bandit problems where the parameter space is factorizable, and it overcomes the limitations of the two major existing approaches by improving both the scalability and robustness. Besides, when combined with the offline-training-online-deployment schedule, `MTSS` yields low system latency and is suitable for large-scale systems. The framework is attractive for cold-start problems as well.

Second, we discuss three concrete examples thoroughly, including ranking, combinatorial optimization, and assortment optimization. These problems have attracted great interest in the literature due to the importance. We provide a novel and practical solution to these application domains.

Third, we provide a general information-theoretic regret bound (Theorem 1) for `MTSS`, which is easy to adapt to different problems that users care about. The bound decomposes into two parts: the price of learning the generalization function and the regret even with the generalization function known in advance. As an example, we derive the regret bound under semi-bandits (Theorem 2) and show that the regret of `MTSS` due to not knowing the generalization function is asymptotically negligible and does not grow with the number of items N, unlike the feature-agnostic approach. Furthermore, the regret of the feature-determined approach scales linearly with the number of time points T, due to its restricted model assumption. These results highlight the benefits of meta-learning.

Finally, in three applications, we compare our approach with existing ones using extensive experiments on both synthetic and real datasets. The results show that the proposed framework can learn efficiently in large problems (Section 7), is computationally attractive (Section 7.1), yields robustness to model misspecification (Appendix F.1), and is useful for cold-start problems (Appendix F.3).

2 SETUP

We consider the following popular and general class of bandit problems [Russo et al., 2017]:

$$\begin{aligned} \boldsymbol{Y}_t &\sim f(\boldsymbol{Y}_t|A_t,\boldsymbol{\theta}),\\ R_t &= f_r(\boldsymbol{Y}_t;\boldsymbol{\eta}). \end{aligned} \tag{1}$$

Here, for $t = 1,\dots,T$, the agent will sequentially choose action A_t from the action space $\mathcal{A}$ and then receive corresponding stochastic observations $\boldsymbol{Y}_t$, which determines the reward R_t through a deterministic function f_r with some *known* parameters $\boldsymbol{\eta}$. The observation $\boldsymbol{Y}_t$ is generated following a domain model f with some *unknown* parameters $\boldsymbol{\theta}$. In many real problems, f is typically a complex distribution involving nonlinear functions, $\boldsymbol{\theta}$ is high-dimensional, and the action space $\mathcal{A}$ is huge. Denote $r(a,\boldsymbol{\theta}) = \mathbb{E}(R_t|A_t = a,\boldsymbol{\theta})$ as the expected reward of taking action a in a problem instance with parameter $\boldsymbol{\theta}$. One common metric is the cumulative regret

$$R(T,\boldsymbol{\theta}) = \sum\nolimits_{t=1}^{T}\Big[\max_{a\in\mathcal{A}} r(a,\boldsymbol{\theta}) - r(A_t,\boldsymbol{\theta})\Big].$$

In many applications, the structured bandit problem consists of N items, and the unknown parameter $\boldsymbol{\theta}$, admittedly being high-dimensional, can be factorized over these items as $\boldsymbol{\theta} = (\theta_1,\dots,\theta_N)^T$, where θ_i is the parameter related to the ith item. This problem setting subsumes many popular bandit problems, such as dynamic assortment optimization where the agent needs to recommend a subset of items, online learning to rank where the agent needs to generate a ranked slate, combinatorial semi-bandits which have numerous applications including online advertisement and optimal network routing, and many others. In this paper, we will focus on this class of structured bandit problems, and will discuss three representative examples in Section 5.

3 LIMITATION OF EXISTING APPROACHES

The existing works typically study one specific task in this class, and as discussed in Section 1, two major approaches dominate this area in the past decade: the *feature-agnostic* approach and the *feature-determined* approach. Feature-agnostic methods [Chen et al., 2013, Wang and Chen, 2018, Kveton et al., 2015, Cheung et al., 2019, Agrawal et al., 2017, 2019] do not utilize side information such as features and learn each θ_i independently. Most of them adapt either the upper-confidence bound (UCB) or TS framework. In these works, the regret bounds will scale quickly with the number of items N, which could be prohibitive in many modern applications. Therefore, feature-agnostic approaches are known to be (statistically) *non-scalable*, and in some experiments, even show a (nearly) linear regret [Wen et al., 2015, Zong et al., 2016, Ou et al., 2018, Agrawal et al., 2020].

To address the scalability issue, feature-determined approaches [Wen et al., 2015, Zong et al., 2016, Ou et al., 2018, Agrawal et al., 2020] utilize the feature vector $\mathbf{x}_i$ of each item i, by assuming a *deterministic* function g parameterized by $\boldsymbol{\gamma}$ such that $\theta_i = g(\mathbf{x}_i;\boldsymbol{\gamma})$ with no error. Under this generalization model assumption, the regret bound for feature-determined approaches can be independent of

Table 1: Comparison of key model assumptions.

Feature-agnostic	Feature-determined	Feature-guided (ours)
$\theta_i \sim \mathbb{P}(\theta)$	$\theta_i = g(\mathbf{x}_i; \boldsymbol{\gamma})$	$\theta_i \sim g(\theta_i \vert \mathbf{x}_i, \boldsymbol{\gamma})$

N but depend on the number of features d instead, which is a theoretically attractive argument when $d = o(N)$.

However, feature-determined approaches have two major limitations. First, feature-determined approaches are typically *computationally demanding* for online updating, which may cause system latency issues in online deployment. This is due to that we add an additional layer to the already complex structured bandit model and also have to update the full model as a whole. Second and more seriously, as usual, algorithms designed with a restrictive model assumption are brittle. No matter how informative $\mathbf{x}_i$ is and how complex g is, it is typically challenging to ensure $\theta_i \equiv g(\mathbf{x}_i)$ *without any error*. To further illustrate, consider a supervised learning task to predict θ_i using $\mathbf{x}_i$. It is hard to believe that there exists a perfect model without any prediction errors. This issue is exacerbated when almost all existing works assume g as linear, given the computational challenge. When this model assumption is violated, the regret is easy to scale linearly with T due to the bias, which is also observed in our experiments. As such, we regard the feature-determined approach as *non-robust* and aim to relax the restricted model assumption.

4 GENERAL FRAMEWORK

To combine the merits of both approaches and hence enable scalable and robust bandit learning, we propose a meta-learning framework, with the model and the algorithm introduced in Section 4.1 and Section 4.2, respectively. In this section, we will focus on the general framework, with examples given in Section 5. For any positive integer M, we denote the set $\{1, \dots, M\}$ by $[M]$.

4.1 Feature-Based Hierarchical Model For Information Sharing

With a large number of items, we adopt the *meta-learning* viewpoint [Vilalta and Drissi, 2002], by regarding the items $\{(\mathbf{x}_i, \theta_i)\}$ as sampled from a joint distribution. To allow information sharing while mitigating the issue from a deterministic generalization model, we model the item-specific parameter θ_i as sampled from a certain distribution $g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma})$ instead of being entirely determined by $\mathbf{x}_i$. Here, g is a model parameterized by an *unknown* vector $\boldsymbol{\gamma}$, which we will instantiate shortly with examples. Therefore, combining with the base model (1), we consider the following hierarchical model:

$$\begin{aligned}
&\text{(Prior)} && \boldsymbol{\gamma} \sim Q(\boldsymbol{\gamma}),\\
&\text{(Generalization function)} && \theta_i|\mathbf{x}_i, \boldsymbol{\gamma} \sim g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma}), \forall i \in [N],\\
&\text{(Observations)} && \boldsymbol{Y}_t \sim f(\boldsymbol{Y}_t|A_t, \boldsymbol{\theta}),\\
&\text{(Reward)} && R_t = f_r(\boldsymbol{Y}_t; \boldsymbol{\eta}),
\end{aligned} \tag{2}$$

where $Q(\boldsymbol{\gamma})$ is the prior distribution for $\boldsymbol{\gamma}$. Intuitively, as such, we can share information across items via g to infer any θ_i and speed up learning, while we can also utilize the observations $\{\boldsymbol{Y}_t\}$ to estimate θ_i in an unbiased way via f. Compared with the two existing approaches, the main difference can be concisely summarized in Table 1.

From the meta-learning perspective, it is more common to consider the Bayes regret [Kveton et al., 2021]:

$$BR(T) = \mathbb{E}_{\boldsymbol{\gamma} \sim Q(\boldsymbol{\gamma}), \theta_i \sim g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma})} R(T, \boldsymbol{\theta}),$$

where the expectation is additionally taken over the item distribution and the prior $Q(\boldsymbol{\gamma})$.

4.2 Meta TS With Feature-Guided Exploration

On the foundation of the hierarchical model (2), we propose Algorithm 1, which is a natural and general TS-type algorithm. TS is one of the most popular bandit algorithm frameworks [Russo et al., 2017, Lattimore and Szepesvári, 2020], with superior numerical and theoretical performance. As a Bayesian algorithm, TS samples the action at each round from the posterior distribution of the optimal action. For a given structured bandit problem, once the generalization model g and the prior are specified, the remaining steps to adapt Algorithm 1 are updating the posterior (steps 1-4) and solving the optimization problem (step 5). This optimization step is problem-dependent, and can typically be solved efficiently via existing methods in the corresponding structured bandit literature.

We will discuss the posterior updating step in depth in Section 4.3. Before we proceed, we remark that step 1-4 of Algorithm 1 can actually be written concisely as sampling $\tilde{\boldsymbol{\theta}}$ from its posterior based on the hierarchical model (2), which can be seen from the relationship

$$\mathbb{P}(\boldsymbol{\theta} \mid \mathcal{H}) = \int_{\boldsymbol{\gamma}} \mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \mathcal{H}) \mathbb{P}(\boldsymbol{\gamma} \mid \mathcal{H}) d\boldsymbol{\gamma}.$$

Therefore, Algorithm 1 can be regarded as a TS-type algorithm. We split the posterior updating process into steps 1-4 for two major reasons. First, in many cases, it is computationally more efficient to update the posteriors of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ separately, as will be discussed in the next section. Second, this decomposition provides a nice insight that our framework actually constructs a feature-based informative prior $g(\theta_i|\mathbf{x}_i, \tilde{\boldsymbol{\gamma}})$ for each θ_i to guide the feature-agnostic TS algorithm, and the prior is obtained by pooling infor-

Algorithm 1: MTSS: Meta Thompson Sampling for Structured bandits

```
Input : Prior $Q(\boldsymbol{\gamma})$ and known parameters of the
        model
Set $\mathcal{H}_1 = \{\}$
while $t < T$ do
    1. Update the posterior of $\boldsymbol{\gamma}$ as $\mathbb{P}(\boldsymbol{\gamma}|\mathcal{H}_t)$, according
       to the hierarchical model (2)
    2. Sample $\tilde{\boldsymbol{\gamma}} \sim \mathbb{P}(\boldsymbol{\gamma}|\mathcal{H}_t)$
    3. Update the posterior of $\boldsymbol{\theta}$ as $\mathbb{P}(\boldsymbol{\theta}|\mathcal{H}_t, \tilde{\boldsymbol{\gamma}})$,
       according to model (1) with $g(\theta_i \mid \mathbf{x}_i, \tilde{\boldsymbol{\gamma}})$ as the
       prior for each $\theta_i$
    4. Sample $\tilde{\boldsymbol{\theta}} \sim \mathbb{P}(\boldsymbol{\theta}|\mathcal{H}_t, \tilde{\boldsymbol{\gamma}})$
    5. Take the greedy action $A_t$ w.r.t. $\tilde{\boldsymbol{\theta}}$ as
       $A_t = \arg\max_{a \in \mathcal{A}} \mathbb{E}(R_t \mid a, \tilde{\boldsymbol{\theta}})$
    6. Receive reward $R_t$ and update the dataset as
       $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{(A_t, R_t)\}$
end
```

mation across items via their features using the hierarchical model. As such, our approach is an instance of meta-learning [Vilalta and Drissi, 2002], and hence we refer to Algorithm 1 as Meta Thompson Sampling for Structured bandits (MTSS).

Remark 1. *The proposed framework is particularly useful for cold-start problems, where new items will be frequently added. Without any historical interaction data, it is important to construct an informative prior for a new item based on its features to guide the exploration.*

4.3 Posterior Updating And Offline-Training-Online-Deployment

In Algorithm 1, the posterior updating can be computed either explicitly when the problem structure permits (see e.g., Section 5.2), or via approximate posterior inference algorithms, such as Gibbs sampler [Johnson et al., 2010] or variational inference [Blei et al., 2017]. We note that the base model (1) typically yields a nice conjugate structure for $\boldsymbol{\theta}$ (e.g., in all three examples in Section 5), and approximate posterior inference can be applied to $\boldsymbol{\gamma}$ alone in these cases. Approximate posterior inference is widely applied to TS [Yu et al., 2020, Wan et al., 2021], and is particularly appropriate in this case due to two reasons: (i) the posterior of $\boldsymbol{\gamma}$ is only used to construct a prior for the base model (1), and hence its error will not be destructive, as related feature-agnostic TS algorithms typically enjoy prior-independent or instance-independent sublinear regrets [Wang and Chen, 2018, Perrault et al., 2020, Zhong et al., 2021] (see Appendix F.1 for details); (ii) when computing the posterior of $\boldsymbol{\gamma}$, many approximate inference algorithms can benefit from the hierarchical structure and hence be efficient. For example, with Gibbs sampler, the algorithm will alternate between the posterior of $\boldsymbol{\theta}$, which typically yields a conjugate form, and that of $\boldsymbol{\gamma}$, which involves a Bayesian regression. Both parts can be solved efficiently.

To facilitate computationally efficient deployment, we further propose an offline-training-online-deployment variant, where we only sample a new $\tilde{\boldsymbol{\gamma}}$ at a certain time point $t \in \mathcal{T}$ instead of at every time point. For example, $\mathcal{T}$ can be $\{2^l : l = 1, 2, \dots\}$ or some trigger time every week. In other words, we will re-train the generalization model $g(\theta; \mathbf{x}, \boldsymbol{\gamma})$ *offline* in a batch mode, and utilize the priors $\{g(\theta_i|\mathbf{x}_i, \tilde{\boldsymbol{\gamma}})\}$ during *online* deployment. As such, during the online phase, our algorithm requires *zero* additional computational cost compared to feature-agnostic TS. Therefore, MTSS in general yields low latency and hence is suitable for large-scale systems. Besides, a powerful generalization function such as a Gaussian process or a Bayesian neural network also becomes feasible. This is a highly practical algorithm, and our numerical results further support its good performance. Finally, it can also be viewed as an empirical Bayes approach [Maritz and Lwin, 2018].

5 EXAMPLES

In this section, we illustrate our framework with three representative examples. For every example, we will first write its feature-agnostic form as model (1), then discuss its applications and the optimization problem, next instantiate model (2) with an *example* choice of g, and finally discuss the corresponding posterior computation to instantiate MTSS. Denote the cardinality of set A by $|A|$.

5.1 Cascading Bandits For Online Learning To Rank

The cascading model is popular in learning to rank [Chuklin et al., 2015] to characterize how a user interacts with an ordered list of K items. Its bandit version has attracted much attention recently, and both feature-agnostic [Kveton et al., 2015, Cheung et al., 2019] and feature-determined approaches [Zong et al., 2016] have been discussed. In this model, $\mathcal{A}$ contains all the subsets of length K, $A_t = (a_t^1, \dots, a_t^K) \in \mathcal{A}$ is a sorted list of items being displayed, $\boldsymbol{Y}_t$ is an indicator vector with the ath entry equal to 1 when the ath displayed item is clicked, and R_t is the reward with $f_r(\boldsymbol{Y}_t) \equiv \sum_{k \in [K]} Y_{k,t} \in \{0, 1\}$, where $Y_{k,t}$ is the kth entry of $\boldsymbol{Y}_t$. The model is intuitive and widely applied: the user will exam the K displayed items from top to bottom, and stop to click one item once she is attracted (or leave if none of them is attractive). Let I_t be the index of the chosen item if exists, and otherwise let $I_t = K$. To formally define the model f, it is useful to introduce a latent binary variable $E_{k,t}$ to indicate if the kth displayed item is examined by the tth user, and a latent variable $W_{k,t}$ to indicate if the kth displayed item is attractive to the tth user. Therefore, the value of $W_{k,t}$ is only visible when $k \leq I_t$. Let θ_i be the attractiveness of the item i. The key probabilistic assumption is that $W_{k,t} \sim Bernoulli(\theta_{a_t^k}), \forall k \in [K]$. When $\boldsymbol{\theta}$ is known, the optimal action is any permutation of the top K items with the highest attractiveness factors.

To characterize the relationship between items using their features, one example choice of g is the popular Beta-Bernoulli logistic model [Forcina and Franconi, 1988, Wan et al., 2021], where $\theta_i {\sim} Beta(logistic(\mathbf{x}_i^T\boldsymbol{\gamma}), \phi)$ for some known parameter ϕ. Hereinafter, we adopt the mean-precision parameterization of the Beta distribution, with $logistic(\mathbf{x}_i^T\boldsymbol{\gamma})$ being the mean and ϕ being the precision parameter. Therefore, our model is

$$\begin{aligned}
\theta_i &\sim Beta(logistic(\mathbf{x}_i^T\boldsymbol{\gamma}), \phi), \forall i \in [N],\\
Y_{k,t} &= W_{k,t}E_{k,t}, \forall k \in [K],\\
W_{k,t} &\sim Bernoulli(\theta_{a_t^k}), \forall k \in [K],\\
E_{k,t} &= (1 - Y_{k-1})E_{k-1,t}, \forall k \in [K],\\
R_t &= \sum\nolimits_{k\in[K]} Y_{k,t},
\end{aligned}$$

with $E_{1,t} \equiv 1$. With a given $\boldsymbol{\gamma}$, the posterior of $\boldsymbol{\theta}$ enjoys the Beta-Bernoulli conjugate relationship and hence can be updated explicitly and efficiently. The prior $Q(\boldsymbol{\gamma})$ can be chosen as many appropriate distributions such as Gaussian. To update the posterior of $\boldsymbol{\gamma}$, we can apply approximate inference as discussed in Section 4.3. Many other learning to rank models, such as the position-based model [Chuklin et al., 2015], can be formulated and solved similarly.

5.2 Combinatorial Semi-Bandits For Online Combinatorial Optimization

Online combinatorial optimization has numerous applications [Sankararaman, 2016], including maximum weighted matching, ads allocation, webpage optimization, etc. It is common that every chosen item will generate a separate observation, known as the semi-bandit problem [Chen et al., 2013]. Both the feature-agnostic [Chen et al., 2013, Wang and Chen, 2018] and the feature-determined [Wen et al., 2015] approaches have been studied. Formally, in a combinatorial semi-bandit, the feasible set $\mathcal{A} \subseteq \{A \subseteq [N] : |A| \leq K\}$ consists of subsets that satisfy the size constraint and other application-specific constraints. The agent will sequentially choose a subset A_t, and then receive a reward $Y_{i,t}$ for each chosen item $i \in A_t$. The overall reward is $R_t = \sum_{i\in A_t} Y_{i,t}$. With known mean rewards, the optimal action can be obtained from a combinatorial optimization problem, which can be efficiently solved in most real applications considered in the semi-bandit literature [Chen et al., 2013]. As an example, we focus on the popular case where $Y_{i,t}$ is Gaussian and consider using a linear mixed model (LMM) as the generalization model. Specifically, the full model is

$$\begin{aligned}
\theta_i &\sim \mathcal{N}(\mathbf{x}_i^T\boldsymbol{\gamma}, \sigma_1^2), \forall i \in [N],\\
Y_{i,t} &\sim \mathcal{N}(\theta_i, \sigma_2^2), \forall i \in A_t,\\
R_t &= \sum\nolimits_{i\in A_t} Y_{i,t},
\end{aligned} \tag{3}$$

where it is typically assumed that σ_1 and σ_2 are known. We choose the prior $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ with parameters as known. For this instance, the posteriors can be derived explicitly (see Appendix B). Many other distributions (e.g., Bernoulli) and model assumptions (e.g., Gaussian process) can be formulated similarly, depending on the applications.

5.3 MNL Bandits For Dynamic Assortment Optimization

Assortment optimization [Pentico, 2008] is a long-standing problem that aims to offer the most profitable subset of items, especially when there exist substitution effects. The Multinomial Logit (MNL) model [Luce, 2012] is arguably the most popular one, and the corresponding bandit problem has been studied, via either the feature-agnostic [Agrawal et al., 2017, 2019] or the feature-determined approaches [Ou et al., 2018, Agrawal et al., 2020]. In assortment optimization, the agent offers a subset (assortment) $A_t \in \mathcal{A} = \{A \subseteq [N] : |A| \leq K\}$, then the customer will choose either one of them or the no-purchase option (denoted as item 0). Let $\boldsymbol{Y}_t = (Y_{0,t}, \cdots, Y_{N,t})^T$ be an indicator vector, where $Y_{i,t} = 1$ if the item i is chosen. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)^T$, where η_k is the revenue of the item k. The reward in round t is then $R_t = \sum_{i\in A_t} Y_{i,t}\eta_i$. In an MNL bandit, each item i has an utility factor v_i, and the choice behaviour follows

$$\boldsymbol{Y}_t \sim Multinomial(1, \frac{v_i \mathbb{I}(i \in \{0\} \cup A_t)}{1 + \sum_{j\in A_t} v_j}),$$

with the convention that $v_0 = 1$. When v_i's are known, the optimal assortment can be solved via linear programming [Agrawal et al., 2017].

Since direct inference under this model is intractable due to the complex dependency of the reward distribution on A_t, an epoch-type offering [Agrawal et al., 2017, 2019, Dong et al., 2020] is more popular in the bandit literature, where we keep offering the same assortment A^l in the lth epoch until the no-purchase appears. Under this setup, it is easier to work with the item-specific parameter $\theta_i = (1 + v_i)^{-1}$ and consider the number of purchase for the item i in each epoch, denoted as Y_i^l. Then, based on Lemma 1 in Agrawal et al. [2017], it can be proven that $Y_i^l \sim Geometric(\theta_i), \forall i \in A^l$. The nice property of such a schedule is that the distributions do not depend on A_t any longer. Besides, the geometric distribution has a nice conjugate relationship with the Beta distribution. As a concrete example of our framework, we can consider modeling

Algorithm 2: MTSS with Epoch-Type Schedule for MNL Bandits

```
Input : Prior $\mathbb{P}(\boldsymbol{\gamma})$ and known parameters of the
        hierarchical model
Set $\mathcal{H}_1 = \{\}$, $t$=1, and $l$=1 keeps track of the time
  steps and total number of epochs, respectively.
while $t < T$ do
    Compute the posterior distribution $\mathbb{P}(\boldsymbol{\theta}|\mathcal{H}_l)$
    For each item $i = 1, \cdots, N$, sample $\tilde{\boldsymbol{\theta}}$ from
      $\mathbb{P}(\boldsymbol{\theta}|\mathcal{H}_l)$, and compute the utility $\tilde{v}_i = \frac{1}{\tilde{\theta}_i} - 1$
    Compute $A^l = \arg\max_{a \in \mathcal{A}} \mathbb{E}(R_t \mid a, \tilde{\boldsymbol{\theta}})$;
    while $c_t \neq 0$ do
        Offer $A^l$, observe the purchasing decision $c_t$ of
          the consumer
        Update $\xi_l = \xi_l \cup t$, time indices corresponding
          to epoch $l$
        $t = t + 1$
    end
    For each item $i \in A^l$, compute
      $Y_i^l = \sum_{t \in \xi_l} I(c_t = i)$, which is the number of
      picks of item i in epoch $l$
    Update the dataset as $\mathcal{H}_{l+1} \leftarrow \mathcal{H}_l \cup \{(A^l, \{Y_i^l\})\}$
    $l = l + 1$
end
```

the relationship between θ_i and $\mathbf{x}_i$ with the following Beta-Geometric logistic model:

$$\theta_i \sim Beta\Big(\frac{logistic(\mathbf{x}_i^T\boldsymbol{\gamma}) + 1}{2}, \phi\Big), \forall i \in [N],$$
$$Y_i^l \sim Geometric(\theta_i), \forall i \in A^l,$$
$$R_l = \sum\nolimits_{i \in A^l} Y_i^l \eta_i.$$

Other generalization models are also possible. We choose this specific form as it is widely observed [Agrawal et al., 2017, 2019] that $v_i < 1$, i.e., the no-purchase option is most popular. This is equal to $\theta_i \in (1/2, 1)$. Finally, we remark that Algorithm 1 needs to be slightly modified to be consistent with the epoch-style offering, though the main idea remains exactly the same. We present the modified MTSS in Algorithm 2. The only difference is that our schedule of sampling new parameters is adjusted to be consistent with the epoch-style. The choices of priors and the posterior updating rules are similar to Section 5.1.

6 THEORY

In this section, we provide theoretical guarantees for MTSS. We start with a general result that provides intuitive insight into the performance of MTSS and is easy to adapt to different specific problems. Our result is information-theoretic and the proof is inspired by Lu and Van Roy [2019]. Let $I(X;Y)$ be the *mutual information* (MI, Kullback [1997]) between two random variables X and Y, $I(X;Y|Z)$ be the conditional MI conditioned on Z, and $I_t(X;Y) = I(X;Y|\{(A'_t, \mathbf{Y}'_t)\}_{t'=1}^{t-1})$. To save space, we defer the detailed definitions to Appendix C. Intuitively, MI measures the mutual dependence between two variables.

Let $\Delta_t = \max_a r(a, \boldsymbol{\theta}) - r(A_t, \boldsymbol{\theta})$ be the per-round regret, and $\mathbb{E}_t(X) = \mathbb{E}(X|\{(A'_t, \mathbf{Y}'_t)\}_{t'=1}^{t-1})$. For a given problem, we assume we can first find some Γ_t and ϵ_t, such that

$$\mathbb{E}_t[\Delta_t] \leq \Gamma_t \sqrt{I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t)} + \epsilon_t, \forall t \in [T]. \tag{4}$$

Here, Γ_t is related to the concentration property of the model, and ϵ_t is a small error term. They can typically be derived by following a few routines introduced in Lu and Van Roy [2019]. We will give an example shortly.

To gain more insights of our bound below, we introduce *oracle-TS*, the TS algorithm that has access to the true generalization model *a priori* and uses $\{g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma})\}$ as priors in feature-agnostic TS. For a general structured bandit problem, the regret of MTSS can be bounded as follows.

Theorem 1. *Suppose that* (4) *holds and* $\Gamma_t \leq \Gamma$ *almost surely for some* Γ. *Then for* MTSS, *we have*

$$BR(T) \leq \underbrace{\Gamma \sum\nolimits_{t=1}^{T} \mathbb{E}[\sqrt{I_t(\boldsymbol{\gamma}; A_t, \mathbf{Y}_t)}]}_{\textit{Regret due to not knowing } \boldsymbol{\gamma}} \tag{5}$$

$$+ \underbrace{\sum\nolimits_{t=1}^{T} \Gamma \mathbb{E}[\sqrt{I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t|\boldsymbol{\gamma})}] + \mathbb{E}[\epsilon_t]}_{\textit{Regret bound for Oracle-TS}}. \tag{6}$$

This decomposition is consistent with our construction, as MTSS aims to learn the generalization model to perform closer to oracle-TS while minimizing the regret. The specific regret bound is problem-dependent. It depends on both the first part of (5) which measures the cost of learning the parameter $\boldsymbol{\gamma}$ (or equivalently, learning the true prior for $\boldsymbol{\theta}$), and the second part which quantifies the unavoidable regret even knowing $\boldsymbol{\gamma}$ (i.e., the performance of oracle-TS). Bounding the first term relies on the MI between the history and $\boldsymbol{\gamma}$, which mainly depends on the concentration property of the hierarchical model. For problems with existing Bayes regret bounds for feature-agnostic TS, the second part can be derived with minimal modifications. The proof of this theorem is differed to Appendix D.1.

As a concrete example, we next analyze the combinatorial semi-bandits with the linear mixed model (see Section 5.2). The results demonstrate the benefits of meta-learning clearly. Without loss of generality, we first state several standard regularity conditions [Basu et al., 2021, Wen et al., 2015, Agrawal et al., 2020, Zhou et al., 2017].

Assumption 1. $\|\mathbf{x}_i\|_2 \leq 1$, for all $i \in [N]$.

Assumption 2. The maximum eigenvalue of $\Sigma_{\boldsymbol{\gamma}}$, $\lambda_1(\Sigma_{\boldsymbol{\gamma}})$, is bounded.

Let $\tilde{O}$ be the big-O notation that hides logarithmic terms, and c_0 be $\lambda_1(\Sigma_{\boldsymbol{\gamma}}) + \sigma_1^2$. We have the following regret bound, the proof of which can be found in Appendix D.2.

Theorem 2. *Under Assumptions 1-2, the Bayes regret of the* `MTSS` *under model* (3) *is bounded by*

$$BR(T) \leq \underbrace{c_1 K\sqrt{Td}\sqrt{log\Big(1+\frac{N\lambda_1(\Sigma_{\boldsymbol{\gamma}})}{\sigma_1^2+\sigma_2^2/T}\Big)}}_{\textit{Regret due to not knowing } \boldsymbol{\gamma}}$$

$$+\underbrace{c_1\sqrt{NTK}\sqrt{log(1+\frac{\sigma_1^2}{\sigma_2^2}T)+K\sqrt{\frac{2}{N}(\lambda_1(\Sigma_{\boldsymbol{\gamma}})+\sigma_1^2)}}}_{\textit{Regret bound for Oracle-TS}}$$

$$= \tilde{O}\Big(K\sqrt{Td}+\sqrt{NTK}\Big),$$

where $c_1 = \sqrt{8log(4NT^2)c_0/log(1+\frac{c_0}{\sigma_2^2})}$.

Therefore, with a large number of items (i.e., $Kd = o(N)$), the regret due to not knowing $\boldsymbol{\gamma}$ is asymptotically negligible (i.e., dominated by the second part), and the performance of `MTSS` is close to oracle-TS, as also observed in experiments. Moreover, note that the second part of the bound is dominated by $c_1\sqrt{NTK}\sqrt{log(1+T\sigma_1^2/\sigma_2^2)}$, which will decay to zero as σ_1 decreases, i.e., when the features become more useful. Therefore, we claim `MTSS` as scalable, since it allows utilizing feature information to learn shared structure so as to behave close to oracle-TS, which yields low regret when the features are informative and serves as the skyline. In contrast, as derived in Basu et al. [2021], the additional regret of feature-agnostic TS than that of oracle-TS can only be bounded by $\sqrt{NTK}$. The dependency on N is as expected, since feature-agnostic TS fails to share information across items and has to learn each from scratch. As such, `MTSS` will be more efficient when features are informative and the number of items is sufficient to learn a good generalization model ($Kd = o(N)$). On the other hand, similar to the discussions in Foster et al. [2020] and Krishnamurthy et al. [2021], feature-determined TS might suffer from the bias as it assumes a restricted model. To our knowledge, as long as $\sigma_1 > 0$, one can only expect a regret bound that is linear in T, which is consistent with our observations in experiments.

7 EXPERIMENTS

7.1 Synthetic Datasets

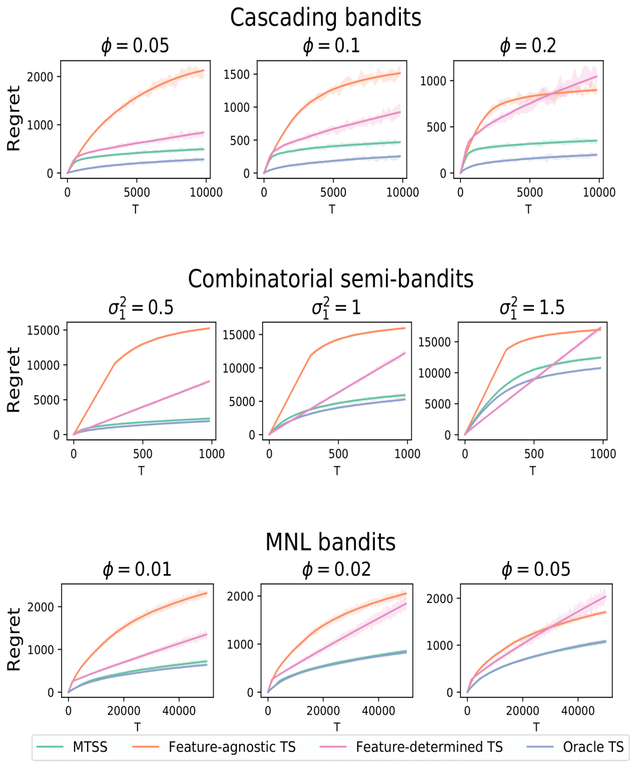


Figure 1: Simulation results. Shaded areas indicate the standard errors of the averages.

Setting. We first conduct simulation experiments to support our theoretical results and investigate the empirical performance of different approaches under various situations. We use the three models introduced in Section 5 to generate data, with (N, K) set as $(1000, 3)$, $(3000, 10)$, $(1000, 5)$ for cascading bandits, semi-bandits, and MNL bandits, respectively. We set $d = 5$ for all tasks, $\boldsymbol{\eta} = \mathbf{1}$ for MNL bandits, and $\sigma_2 = 1$ for semi-bandits. We choose $Q(\boldsymbol{\gamma}) = \mathcal{N}(\mathbf{0}, d^{-1}\boldsymbol{I})$ and sample $\mathbf{x}_i$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ with an intercept. For each problem, we vary the value of either σ_1 or ϕ, where a higher value of σ_1 or ϕ implies a larger heterogeneity between items, conditional on their features.

Baselines. For these three problems, we compare our approach with existing ones, which can be categorized as either feature-agnostic or feature-determined, as we introduced. For the feature-agnostic approaches, we directly apply the TS algorithms proposed in the corresponding papers [Kveton et al., 2015, Wang and Chen, 2018, Agrawal et al., 2017]. For the feature-determined approaches, we compare with the TS algorithms proposed in the corresponding papers [Wen et al., 2015, Zong et al., 2016, Ou et al., 2018]. For fair comparison, we closely follow the spirits of Zong et al. [2016] and Ou et al. [2018], and modify the linear models therein by logistic models to avoid mis-specification in our simulation. We also present the performance of oracle-TS as our skyline. Finally, to study the performance of our algorithm with the offline-training-online-deployment schedule as in Section 4.3, we sample a new $\tilde{\boldsymbol{\gamma}}$ every 500 time points in MNL bandits and cascading bandits, and every 100 time points in semi-bandits.

Results. The experiment results over 50 random seeds are presented in Figure 1. Overall, `MTSS` performs favorably and demonstrates its universality. Our findings can be summarized as follows. First, MTSS enjoys a sublinear regret, while the feature-determined approach suffers from a linear regret due to the bias. This bias becomes more severe when σ_1 or ϕ increases, which implies that the amount of variation in θ_i that can not be explained by $g(\mathbf{x}_i)$ grows. Second, although in general the feature-agnostic methods have a sublinear regret, the learning speed is slow, and hence the cumulative regret is much larger. This is due to the lack of generalization across items. With the offline-training-online-deployment schedule, our algorithm still performs well and is close to oracle-TS. Finally, MTSS is computationally efficient during online updating. For example, on a machine with 96 cores and 192GB RAM, for MNL bandits, the *total* online time costs for feature-agnostic TS and MTSS are 2.3 and 2.5 seconds, respectively. Besides, for all three tasks, even with one million items, the *per-round* cost of MTSS is still less than 0.2 seconds and close to that of feature-agnostic TS, while feature-determined TS is infeasible to finish the task in a reasonable time frame.

Additional Experiments. First, we repeat the experiments with other values of L, K, d in Appendix F.4. The findings are similar. Second, we empirically study the impact of model misspecification in Appendix F.1, where MTSS shows great robustness. Recall that, to facilitate scalability, we assume that $\theta_i|\mathbf{x}_i, \boldsymbol{\gamma} \sim g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma})$. Intuitively, since g is used to construct priors for feature-agnostic TS, our framework is still valuable as long as the learned priors provide reasonable information compared with manually specified ones. This robustness is also supported by the prior-independent or instance-independent sublinear regret bounds for feature-agnostic TS [Wang and Chen, 2018, Perrault et al., 2020, Zhong et al., 2021].

Third, when only a few features are useful (i.e., $\boldsymbol{\gamma}$ is sparse), we demonstrate in Appendix F.2 that the spike-and-slab prior can be used to leverage our framework and enable faster learning. Specifically, let $\boldsymbol{z}$ be a random vector, each entry of which has a Bernoulli prior with probability $\boldsymbol{p}_{slab}$. Given $\boldsymbol{z}$, each entry of $\boldsymbol{\gamma}$ (i.e., γ_i) will be either 0 if $z_i = 0$ or sampled from a distribution Q' if $z_i = 1$. They jointly form our prior distribution $Q(\boldsymbol{\gamma})$, as follows.

$$\begin{aligned}
&\text{(Spike-and-Slab Prior)} & \boldsymbol{\gamma} &\sim \boldsymbol{z} * Q'(\boldsymbol{\gamma}), \\
& & &\text{with } \boldsymbol{z} \sim \text{Bernoulli}(\boldsymbol{p}_{slab}), \\
&\text{(Generalization function)} & \theta_i|\mathbf{x}_i, \boldsymbol{\gamma} &\sim g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma}), \forall i \in [N], \\
&\text{(Observations)} & \boldsymbol{Y}_t &\sim f(\boldsymbol{Y}_t|A_t, \boldsymbol{\theta}), \\
&\text{(Reward)} & R_t &= f_r(\boldsymbol{Y}_t; \boldsymbol{\eta}).
\end{aligned}$$

Finally, in Appendix F.3, we conduct experiments for the cold-start problem, where new items are frequently added and old items are removed. As expected, all algorithms suffer a linear regret in such a changing environment, with MTSS consistently outperforming feature-agnostic TS and feature-determined TS. Specifically, the performance of feature-agnostic TS deteriorates significantly, as no information can be carried over to the new items. On the contrary, the difference between the regret of oracle-TS and MTSS is fairly stable, which implies that MTSS has learned the generalization function well and performs almost the same as oracle-TS eventually.

7.2 Real Data

In this section, we compare MTSS with existing methods (discussed in Section 7.1) on three real datasets. For fair comparisons, we closely follow the related papers to design our experiments. To save space, we describe the main ideas below and defer more details to Appendix F.5.

Datasets. For cascading bandits, we follow Zong et al. [2016] and aim to display ranked restaurants and maximize the probability of the user being attracted to at least one restaurant recommended, using the dataset from Yelp [Asghar, 2016]. In the final dataset, at each round, we display a set of 5 restaurants from a universe of size 3000, and utilize 10 features. For combinatorial semi-bandits, we follow Wen et al. [2015] and aim to send online advertisements to the best subset of users who are most likely to accept the advertisement, while keeping a balance between genders, using the Adult dataset from Dua and Graff [2017]. In the final dataset, we choose a set of 20 users (including ten females and ten males) from a universe of size 3000, and utilize 4 features. For MNL bandits, we follow Oh and Iyengar [2019] and aim to recommend the optimal set of movies, using the MovieLens dataset [Harper and Konstan, 2015]. In the final dataset, we recommend a set of 5 movies from a universe of size 1000, and utilize 5 features.

Design. To simulate data and calculate regrets, we need to first determine $\{\theta_i\}$ and $\{\mathbf{x}_i\}$, and then generate stochastic rewards either by using the base model (1) with $\{\theta_i\}$ as parameters or by directly sampling from the dataset. Again, we closely follow the existing papers. Specifically, for cascading or MNL bandits, we first split the dataset into a training set and a testing set, then estimate the features $\{\mathbf{x}_i\}$ from the training set (via collaborative filtering), and finally estimate the item-specific parameters $\{\theta_i\}$ from the testing set. For semi-bandits, we directly utilize the features and responses in the raw dataset. We remark that, during these procedures, *zero* assumption is imposed manually on the joint distribution of $(\mathbf{x}_i, \theta_i)$ (and hence $\mathbb{P}(\theta_i|\mathbf{x}_i)$), and therefore these setups can be used for fair comparisons between MTSS and the existing approaches.

Results. We present the results in Figure 2. MTSS accumulates lower regrets in all three problems. We observe that feature-agnostic TS suffers the curse of dimensionality and learns slowly. In particular, for cascading or MNL bandits, since the click/purchase rates are low in the two datasets (i.e., useful feedback is sparse), feature-agnostic TS shows a (nearly) linear regret, as also observed in Zong et al.

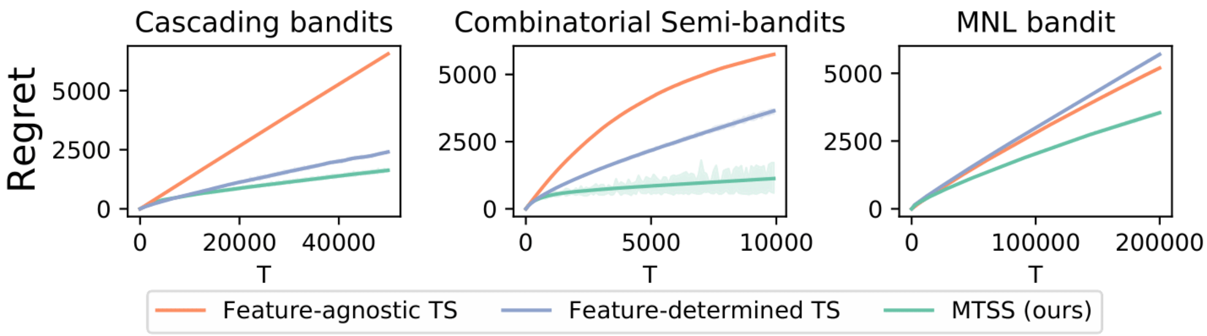


Figure 2: Experiment results on real datasets, averaged over 50 random seeds. Shaded areas indicate the standard errors of the mean, which are small and hence hard to distinguish with some curves.

[2016] and Harper and Konstan [2015]. Besides, while feature-determined TS may slightly outperform in the initial periods, it eventually exhibits a linear trend. This is likely due to the bias from its restrictive model assumption. To provide further support for the proposed method, additional experiments that use larger real datasets and come to similar conclusions have been included in Appendix F.5.4.

8 ADDITIONAL RELATED WORK

Structured Bandits. Standard multi-armed bandits are not scalable to huge action space, and therefore researchers leverage structural information to generalize across actions, known as structured bandits [Van Parys and Golrezaei, 2020]. Besides several stylized models, such as the linear bandits [Chu et al., 2011] and logistic bandits [Kveton et al., 2020], many practical problems depend on domain models and can be summarized as model (1). Besides the two major approaches (i.e., *feature-agnostic* and *feature-determined*) and related papers reviewed in Section 3, Yu et al. [2020] also proposes a framework that unifies a few structured bandit problems. However, this paper mainly focuses on unifying problems without introducing new models and related algorithms for each specific problem as we do. In addition, their approach is restricted to models with only binary variables.

Meta Bandits. Utilizing the framework of hierarchical models, our work is also related to the meta bandits literature [Kveton et al., 2021, Basu et al., 2021, Wan et al., 2021, Hong et al., 2022]. Whereas the focus of meta bandits is on transferring knowledge across a large number of (similar and relatively simple) bandit tasks, such as multi-armed bandits or linear bandits, we focus on information sharing within a single large-scale complicated structured bandit. Therefore, none of the existing methods can be applied to our setup. Furthermore, such a distinction also induces non-trivial differences in the regret analysis, as regrets can no longer be analyzed in separate tasks as meta bandits did. In addition, all existing papers (with the exception of Wan et al. [2021], which is only applicable to multi-armed bandits) only model the tasks as sampled from a simple feature-agnostic distribution, and can not utilize valuable side information such as features in meta-learning as we do. Our novelty lies in proposing a practical and significant meta-learning framework that overcomes the limitations of the two major approaches in the large literature on structured bandits.

9 DISCUSSION

Motivated by the long-standing challenges of learning in large-scale structured bandits, in this paper, we propose a unified meta-learning framework with a TS-type algorithm named `MTSS`. We use three real examples and both numerical and theoretical evidence to demonstrate that the framework is general to subsume a wide class of practical problems, scalable to large systems, and robust to the generalization model assumption.

This approach can be extended in several aspects. First, it is straightforward to allow multiple parameters per item, by fitting one generalization model for each. Second, in our examples, we consider the variance components (σ_1 or ϕ) as known. In practice, we can apply empirical Bayes to update these hyperparameters adaptively (see Appendix A). Third, we mainly focus on TS algorithms as our baselines, since they typically outperform the UCB counterparts and yield fair comparisons with `MTSS`. Although adapting UCB to our setup is not straightforward, Bayesian UCB [Kaufmann et al., 2012] can be similarly developed. Last, it is meaningful to extend our theory to the case with model misspecification or with the offline-training-online-deployment schedule, two variants we empirically studied. This requires a delicate analysis of TS with misspecified priors, which is known as a challenging open problem in the literature [Wan et al., 2021]. We leave these extensions for future research.

References

Agrawal, P., Avadhanula, V., and Tulabandhula, T. (2020). A tractable online learning algorithm for the multinomial logit contextual bandit. *arXiv preprint arXiv:2011.14033*.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2017). Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977*.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2019). Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485.

Asghar, N. (2016). Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Bastani, H., Simchi-Levi, D., and Zhu, R. (2019). Meta dynamic pricing: Transfer learning across experiments. *Available at SSRN 3334629*.

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

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159. PMLR.

Chen, X., Shi, C., Wang, Y., and Zhou, Y. (2021). Dynamic assortment planning under nested logit models. *Production and Operations Management*, 30(1):85–102.

Cheung, W. C., Tan, V., and Zhong, Z. (2019). A thompson sampling algorithm for cascading bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 438–447. PMLR.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Chuklin, A., Markov, I., and Rijke, M. d. (2015). Click models for web search. *Synthesis lectures on information concepts, retrieval, and services*, 7(3):1–115.

Dong, K., Li, Y., Zhang, Q., and Zhou, Y. (2020). Multinomial logit bandit with low switching cost. In *International Conference on Machine Learning*, pages 2607–2615. PMLR.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Ferreira, K. J., Simchi-Levi, D., and Wang, H. (2018). Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602.

Forcina, A. and Franconi, L. (1988). Regression analysis with the beta-binomial distribution. *Rivista di Statistica Applicata*, 21(1).

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. (2020). Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Hong, J., Kveton, B., Zaheer, M., and Ghavamzadeh, M. (2022). Hierarchical bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR.

Johnson, A. A., Jones, G. L., et al. (2010). Gibbs sampling for a bayesian hierarchical general linear model. *Electronic Journal of Statistics*, 4:313–333.

Katariya, S., Kveton, B., Szepesvari, C., Vernade, C., and Wen, Z. (2017). Stochastic rank-1 bandits. In *Artificial Intelligence and Statistics*, pages 392–401. PMLR.

Kaufmann, E., Cappé, O., and Garivier, A. (2012). On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR.

Keshavan, R. H., Montanari, A., and Oh, S. (2009). Low-rank matrix completion with noisy observations: a quantitative comparison. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1216–1222. IEEE.

Krishnamurthy, S. K., Hadad, V., and Athey, S. (2021). Tractable contextual bandits beyond realizability. In *International Conference on Artificial Intelligence and Statistics*, pages 1423–1431. PMLR.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-w., Mladenov, M., Boutilier, C., and Szepesvari, C. (2021). Meta-thompson sampling. *arXiv preprint arXiv:2102.06129*.

Kveton, B., Meshi, O., Zoghi, M., and Qin, Z. (2022). On the value of prior in online learning to rank. In *International Conference on Artificial Intelligence and Statistics*, pages 6880–6892. PMLR.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. (2015). Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776. PMLR.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. (2020). Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Li, S., Wang, B., Zhang, S., and Chen, W. (2016). Contextual combinatorial cascading bandits. In *International conference on machine learning*, pages 1245–1253. PMLR.

Lu, X. and Van Roy, B. (2019). Information-theoretic confidence bounds for reinforcement learning. *arXiv preprint arXiv:1911.09724*.

Luce, R. D. (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Maritz, J. S. and Lwin, T. (2018). *Empirical bayes methods*. Chapman and Hall/CRC.

Oh, M.-h. and Iyengar, G. (2019). Thompson sampling for multinomial logit contextual bandits. *Advances in Neural Information Processing Systems*, 32:3151–3161.

Ou, M., Li, N., Zhu, S., and Jin, R. (2018). Multinomial logit bandit with linear utility functions. *arXiv preprint arXiv:1805.02971*.

Pentico, D. W. (2008). The assortment problem: A survey. *European Journal of Operational Research*, 190(2):295–309.

Perrault, P., Boursier, E., Perchet, V., and Valko, M. (2020). Statistical efficiency of thompson sampling for combinatorial semi-bandits. *arXiv preprint arXiv:2006.06613*.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2017). A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*.

Sankararaman, K. A. (2016). Semi-bandit feedback: A survey of results.

Tomkins, S., Liao, P., Yeung, S., Klasnja, P., and Murphy, S. (2019). Intelligent pooling in thompson sampling for rapid personalization in mobile health.

Van Parys, B. and Golrezaei, N. (2020). Optimal learning for structured bandits. *Available at SSRN 3651397*.

Vilalta, R. and Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95.

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

Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122. PMLR.

Wen, Z., Kveton, B., and Ashkan, A. (2015). Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122. PMLR.

Yu, T., Kveton, B., Wen, Z., Zhang, R., and Mengshoel, O. J. (2020). Graphical models meet bandits: A variational thompson sampling approach. In *International Conference on Machine Learning*, pages 10902–10912. PMLR.

Zhong, Z., Chueng, W. C., and Tan, V. Y. (2021). Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66.

Zhou, Q., Zhang, X., Xu, J., and Liang, B. (2017). Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*, pages 811–821. Springer.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. (2016). Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*.

A Extension: Empirical Bayes for updating variance components adaptively

In our examples, we consider the variance components (σ_1 or ϕ) as known. In practice, we can apply empirical Bayes Maritz and Lwin [2018] to update these hyperparameters adaptively, as in Tomkins et al. [2019] and Wan et al. [2021]. Specifically, suppose the generalization model is $\theta_i|\mathbf{x}_i, \boldsymbol{\gamma} \sim g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma}; \beta)$, where β is a parameter that we assume as known in MTSS. At time point t, given the history $\mathcal{H}_t$, one can focus on the following frequentist model:

$$\begin{aligned} &\text{(Generalization function)} && \theta_i|\mathbf{x}_i, \boldsymbol{\gamma} \sim g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma}; \boldsymbol{\beta}), \forall i \in [N], \\ &\text{(Observations)} && \boldsymbol{Y}_t \sim f(\boldsymbol{Y}_t|A_t, \boldsymbol{\theta}), \\ &\text{(Reward)} && R_t = f_r(\boldsymbol{Y}_t; \boldsymbol{r}). \end{aligned} \tag{7}$$

We write the corresponding likelihood function as $L(\boldsymbol{\theta}, \boldsymbol{\gamma}, \beta|\mathcal{H}_t)$, and let $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\beta})$ be the maximum likelihood estimation. Following the empirical Bayes approach, we use $\theta_i|\mathbf{x}_i, \boldsymbol{\gamma} \sim g(\theta_i|\mathbf{x}_i, \boldsymbol{\gamma}; \hat{\beta})$ in MTSS. The updating of $\hat{\beta}$ can also be periodical.

Intuitively, when the conditional variance decays to 0, our method reduces to feature-determined TS; while when it grows, it indicates the features are less useful, and we are essentially assigning a non-informative prior as commonly adopted in feature-agnostic TS. As such, our framework yields the desired flexibility and is adaptive via empirical Bayes.

B Explicit form of the posterior in semi-bandits with LMM

In this section, we derive the posterior distributions involved in the algorithm for semi-bandits and the proof of Theorem 2. The derivations are standard, and we only include them for completeness.

Recall that $\mathbf{x}_i$ is the features of item i. Let $\boldsymbol{\Phi} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$ contains all N items' features. Let $\boldsymbol{\phi}_t = (\mathbf{x}_k)^T_{k \in A_t}$ be a $|A_t| \times d$ matrix contains features of all items offered at round t, and $\boldsymbol{\Phi}_{1:t} = (\boldsymbol{\phi}_1^T, \cdots, \boldsymbol{\phi}_t^T)^T$ is a $C_t \times d$ matrix including features of all the item offered from round 1 to round t, where $C_t = \sum_{l=1}^t |A_l|$. Likewise, $\boldsymbol{Y}_{1:t} = (\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_t^T)^T$ includes observed rewards of all items offered till round t. Then, we define a $N \times C_t$ matrix $\boldsymbol{Z}_{1:t}$, such that the (j, a)-th entry of $\boldsymbol{Z}_{1:t}$ is $\mathbb{I}(i(a) = j)$, $j \in [N]$. Here, $i(a)$ is the item index of the ath observed reward in $\boldsymbol{Y}_{1:t}$. The row i of $\boldsymbol{Z}_{1:t}$ is defined as $\boldsymbol{Z}_{1:t,i}$. Finally, we define that $n_t(i)$ is the total number of pulls of arm i from round 1 till round t, include round t.

Recall that our model for semi-bandits is defined as following:

$$\begin{aligned} \boldsymbol{\gamma} &\sim \mathcal{N}(\boldsymbol{\mu_\gamma}, \boldsymbol{\Sigma_\gamma}), \\ \theta_i &\sim \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\gamma}, \sigma_1^2), \forall i \in [N], \\ Y_{i,t} &\sim \mathcal{N}(\theta_i, \sigma_2^2), \forall i \in A_t. \end{aligned}$$

Posterior Distribution of $\boldsymbol{\theta}$ Given $\mathcal{H}_{t+1}$:

First, we compute the distribution of $\boldsymbol{Y}_{1:t}$ given $\boldsymbol{\Phi}_{1:t}(\boldsymbol{Z}_{1:t})$ and $\boldsymbol{\theta}$, the distribution of $\boldsymbol{Y}_{1:t}$ given only $\boldsymbol{\Phi}_{1:t}$, and the distribution of $\boldsymbol{\theta}$. Note that, $\boldsymbol{\Phi}_{1:t} = \boldsymbol{Z}_{1:t}^T \boldsymbol{\Phi}$. Given $\boldsymbol{\theta}$, we can write

$$\boldsymbol{Y}_{1:t} = \boldsymbol{Z}_{1:t}^T \boldsymbol{\theta} + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma_2^2 \boldsymbol{I}_{C_t}).$$

Similarly, given $\boldsymbol{\Phi}$ and $\boldsymbol{\gamma}$, we have

$$\boldsymbol{\theta} = \boldsymbol{\Phi}\boldsymbol{\gamma} + v, \text{ where } v \sim \mathcal{N}(0, \sigma_1^2 \boldsymbol{I}_N).$$

Further, given $\boldsymbol{\mu_\gamma}$, we have

$$\boldsymbol{\gamma} = \boldsymbol{\mu_\gamma} + b, \text{ where } b \sim \mathcal{N}(0, \boldsymbol{\Sigma_\gamma}).$$

Combining above three equations, we have

$$\begin{aligned} \boldsymbol{Y}_{1:t}|\boldsymbol{\Phi}_{1:t}, \boldsymbol{\theta} &= \boldsymbol{Z}_{1:t}^T \boldsymbol{\theta} + \epsilon, \\ \boldsymbol{\theta} &= \boldsymbol{\Phi}\boldsymbol{\mu_\gamma} + \boldsymbol{\Phi} b + v, \\ \boldsymbol{Y}_{1:t}|\boldsymbol{\Phi}_{1:t} &= \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu_\gamma} + \boldsymbol{\Phi}_{1:t} b + \boldsymbol{Z}_{1:t}^T v + \epsilon. \end{aligned}$$

Therefore, we have

$$\begin{aligned}
\boldsymbol{Y}_{1:t}|\boldsymbol{\Phi}_{1:t},\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{Z}_{1:t}^T\boldsymbol{\theta}, \sigma_2^2\boldsymbol{I}_{C_t}),\\
\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}^T + \sigma_1^2\boldsymbol{I}_N),\\
\boldsymbol{Y}_{1:t}|\boldsymbol{\Phi}_{1:t} &\sim \mathcal{N}(\boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t}).
\end{aligned}$$

Let $\boldsymbol{B} \sim \mathcal{N}(0, \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}^T + \sigma_1^2\boldsymbol{I}_N)$, we have

$$\begin{aligned}
\boldsymbol{\theta} &= \boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \boldsymbol{B},\\
\boldsymbol{Y}_{1:t}|\boldsymbol{\Phi}_{1:t},\boldsymbol{B} &\sim \mathcal{N}(\boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \boldsymbol{Z}_{1:t}^T\boldsymbol{B}, \sigma_2^2\boldsymbol{I}_{C_t}).
\end{aligned}$$

Then, we compute the posterior distribution of $\boldsymbol{B}$ instead of $\boldsymbol{\theta}$.

$$\begin{aligned}
\mathbb{P}(\boldsymbol{B}|\boldsymbol{Y}_{1:t}) &\propto \mathbb{P}(\boldsymbol{Y}_{1:t}|\boldsymbol{B})\mathbb{P}(\boldsymbol{B})\\
&\propto exp\Big(-\frac{1}{2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}} - \boldsymbol{Z}_{1:t}^T\boldsymbol{B})^T\frac{1}{\sigma_2^2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}} - \boldsymbol{Z}_{1:t}^T\boldsymbol{B})\Big)\\
&\quad \times exp\Big(-\frac{1}{2}\boldsymbol{B}^T(\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}^T + \sigma_1^2\boldsymbol{I}_N)^{-1}\boldsymbol{B}\Big)\\
&\propto exp\Big(\boldsymbol{B}^T\boldsymbol{Z}_{1:t}\frac{1}{\sigma_2^2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}})\\
&\qquad - \frac{1}{2}\boldsymbol{B}^T\underbrace{\{(\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}^T + \sigma_1^2\boldsymbol{I}_N)^{-1} + \frac{1}{\sigma_2^2}\boldsymbol{Z}_{1:t}\boldsymbol{Z}_{1:t}^T\}}_{\tilde{\boldsymbol{\Sigma}}^{-1}}\boldsymbol{B}\Big)\\
&\sim \mathcal{N}(\underbrace{\tilde{\boldsymbol{\Sigma}}\boldsymbol{Z}_{1:t}\frac{1}{\sigma_2^2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}})}_{\boldsymbol{\mu}(\boldsymbol{B})}, \tilde{\boldsymbol{\Sigma}}).
\end{aligned}$$

Using the Woodbury matrix identity [Rasmussen, 2003], we have

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}^T + \sigma_1^2\boldsymbol{I}_N\\
&\quad - (\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t})\Big(\sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t}\Big)^{-1}(\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t})^T,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\mu}(\boldsymbol{B}) &= \tilde{\boldsymbol{\Sigma}}\boldsymbol{Z}_{1:t}\frac{1}{\sigma_2^2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}})\\
&= (\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t})\Big[\boldsymbol{I}_{C_t} - \Big(\sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t}\Big)^{-1}\\
&\qquad\qquad (\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t})^T\boldsymbol{Z}_{1:t}\Big]\frac{1}{\sigma_2^2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}})\\
&= (\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t})\Big(\sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t}\Big)^{-1}\sigma_2^2\boldsymbol{I}_{C_t}\frac{1}{\sigma_2^2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}})\\
&= (\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t})\Big(\sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t}\Big)^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}).
\end{aligned}$$

Since $\boldsymbol{\theta} = \boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \boldsymbol{B}$, we get the posterior distribution of $\boldsymbol{\theta}$.

$$\boldsymbol{\theta}|\mathcal{H}_{t+1} \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \boldsymbol{\mu}(\boldsymbol{B}), \tilde{\boldsymbol{\Sigma}}).$$

In particular, for each item $i \in [N]$, the posterior distribution of the item-specific parameter θ_i is as follows.

$$\begin{aligned}
&\theta_i|\mathcal{H}_{t+1} \sim \mathcal{N}(\hat{\mu}_{t+1}(i), \hat{\sigma}_{t+1}^2(i)),\\
&\hat{\mu}_{t+1}(i) = \mathbf{x}_i^T\boldsymbol{\mu}_{\boldsymbol{\gamma}} + (\mathbf{x}_i\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t,i})\\
&\qquad\qquad \times (\sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}),\\
&\hat{\sigma}_{t+1}^2(i) = \mathbf{x}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\mathbf{x}_i + \sigma_1^2 - (\mathbf{x}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t,i})\\
&\qquad\qquad \times (\sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t})^{-1}(\mathbf{x}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t,i})^T.
\end{aligned} \tag{8}$$

Alternatively, since

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}^{-1} &= (\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}^T + \sigma_1^2\boldsymbol{I}_N)^{-1} + \frac{1}{\sigma_2^2}\boldsymbol{Z}_{1:t}\boldsymbol{Z}_{1:t}^T \\
&= (\boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}^T + \sigma_1^2\boldsymbol{I}_N)^{-1} + \frac{1}{\sigma_2^2}diag(n_t(i))_{i=1}^N,
\end{aligned}$$

then,

$$\hat{\sigma}_{t+1}^{-2}(i) = \hat{\sigma}_t^{-2}(i) + \frac{1}{\sigma_2^2}(n_t(i) - n_{t-1}(i)).$$

Posterior Distribution of $\boldsymbol{\gamma}$ Given $\mathcal{H}_{t+1}$:

Similarly, we can write

$$\boldsymbol{Y}_{1:t}|\boldsymbol{\Phi}_{1:t}, b \sim \mathcal{N}(\boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \boldsymbol{\Phi}_{1:t}b, \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t}).$$

Then we compute the posterior distribution of b instead of $\boldsymbol{\gamma}$.

$$\begin{aligned}
\mathbb{P}(b|\boldsymbol{Y}_{1:t}) &\propto \mathbb{P}(\boldsymbol{Y}_{1:t}|b)\mathbb{P}(b) \\
&\propto exp\Big[-\frac{1}{2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}} - \boldsymbol{\Phi}_{1:t}b)^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1} \\
&\qquad\qquad \times (\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}} - \boldsymbol{\Phi}_{1:t}b)\Big] exp\Big[\frac{1}{2}b^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}b\Big] \\
&\propto exp\Big[-\frac{1}{2}b^T\underbrace{(\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}\boldsymbol{\Phi}_{1:t} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1})}_{\boldsymbol{\Sigma}_*^{-1}}b \\
&\qquad\qquad + b^T\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}})\Big] \\
&\sim \mathcal{N}(\underbrace{\boldsymbol{\Sigma}_*\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}})}_{\boldsymbol{\mu}_*}, \boldsymbol{\Sigma}_*).
\end{aligned}$$

Using the Woodbury matrix identity [Rasmussen, 2003], we have

$$\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T)^{-1}\boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}},$$

and

$$\begin{aligned}
\boldsymbol{\mu}_* &= \boldsymbol{\Sigma}_*\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}) \\
&= \Big(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T)^{-1}\boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\Big)\boldsymbol{\Phi}_{1:t}^T \\
&\qquad \times (\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T\Big[\boldsymbol{I} - (\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T)^{-1}\boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T\Big] \\
&\qquad \times (\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T)^{-1}(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t}) \\
&\qquad \times (\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T)^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}).
\end{aligned}$$

To derive an explicit form of $\boldsymbol{\Sigma}_*$, we focus on $\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}\boldsymbol{\Phi}_{1:t} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}$. Again, using the Woodbury matrix identity, we have

$$\begin{aligned}
(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1} &= \sigma_2^{-2}\boldsymbol{I}_{C_t} - \sigma_2^{-4}\boldsymbol{Z}_{1:t}^T(\sigma_1^{-2}\boldsymbol{I}_N + \sigma_2^{-2}\boldsymbol{Z}_{1:t}\boldsymbol{Z}_{1:t}^T)^{-1}\boldsymbol{Z}_{1:t} \\
&= \sigma_2^{-2}\boldsymbol{I}_{C_t} - \sigma_2^{-4}\boldsymbol{Z}_{1:t}^T diag\Big(\frac{1}{\sigma_1^{-2} + \sigma_2^{-2}n_t(i)}\Big)_{i=1}^N\boldsymbol{Z}_{1:t}.
\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_*^{-1} &= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t})^{-1}\boldsymbol{\Phi}_{1:t}\\ &= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \boldsymbol{\Phi}_{1:t}^T\Big(\sigma_2^{-2}\boldsymbol{I}_{C_t} - \sigma_2^{-4}\boldsymbol{Z}_{1:t}^T\mathcal{D}\boldsymbol{Z}_{1:t}\Big)\boldsymbol{\Phi}_{1:t}\\ &= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \sigma_2^{-2}\boldsymbol{\Phi}_{1:t}^T\boldsymbol{\Phi}_{1:t} - \sigma_2^{-4}\boldsymbol{\Phi}_{1:t}^T\boldsymbol{Z}_{1:t}^T\mathcal{D}\boldsymbol{Z}_{1:t}\boldsymbol{\Phi}_{1:t}\\ &= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \sum_{i=1}^N \frac{n_t(i)}{\sigma_2^2 + \sigma_1^2 n_t(i)}\mathbf{x}_i\mathbf{x}_i^T,\end{aligned}$$

where $\mathcal{D} = diag\Big(\frac{1}{\sigma_1^{-2}+\sigma_2^{-2}n_t(1)}, \cdots, \frac{1}{\sigma_1^{-2}+\sigma_2^{-2}n_t(N)}\Big)$.

Therefore,

$$\begin{aligned}&\boldsymbol{\gamma}|\mathcal{H}_{t+1} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_{t+1}, \tilde{\boldsymbol{\Sigma}}_{t+1}),\\ &\quad\tilde{\mu}_{t+1} = \boldsymbol{\mu}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T(\sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t} + \sigma_2^2\boldsymbol{I}_{C_t} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T)^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\mu}_{\boldsymbol{\gamma}}),\\ &\quad\tilde{\boldsymbol{\Sigma}}_{t+1}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \sum_{i=1}^N \frac{n_t(i)}{\sigma_2^2 + \sigma_1^2 n_t(i)}\mathbf{x}_i\mathbf{x}_i^T.\end{aligned} \tag{9}$$

Posterior Distribution of $\boldsymbol{\theta}$ Given $\mathcal{H}_{t+1}$ and $\boldsymbol{\gamma}$:

Similarly, to derive the posterior distribution $\boldsymbol{\theta}$ given $\mathcal{H}_{t+1}$ and $\boldsymbol{\gamma}$, we first derive the posterior distribution of v given $\mathcal{H}_{t+1}$ and $\boldsymbol{\gamma}$. Here, we can write

$$\boldsymbol{Y}_{1:t}|\boldsymbol{\Phi}_{1:t}, \boldsymbol{\gamma}, v \sim \mathcal{N}(\boldsymbol{\Phi}_{1:t}\boldsymbol{\gamma} + \boldsymbol{Z}_{1:t}^T v, \sigma_2^2\boldsymbol{I}_{C_t}).$$

Then, the posterior distribution of v given $\mathcal{H}_{t+1}$ and $\boldsymbol{\gamma}$ is

$$\begin{aligned}&\mathbb{P}(v|\boldsymbol{Y}_{1:t}, \boldsymbol{\gamma})\\ &\propto \mathbb{P}(\boldsymbol{Y}_{1:t}|v, \boldsymbol{\gamma})\mathbb{P}(v|\boldsymbol{\gamma})\\ &\propto exp\Big(-\frac{1}{2}\sigma_2^{-2}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\gamma} - \boldsymbol{Z}_{1:t}^T v)^T(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\gamma} - \boldsymbol{Z}_{1:t}^T v)\Big)exp\Big(-\frac{1}{2}\sigma_1^{-2}v^T v\Big)\\ &\propto exp\Big(-\frac{1}{2}v^T(\sigma_2^{-2}\boldsymbol{Z}_{1:t}\boldsymbol{Z}_{1:t}^T + \sigma_1^{-2}\boldsymbol{I}_N)v + v^T\sigma_2^{-2}\boldsymbol{Z}_{1:t}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\gamma})\Big)\\ &\sim \mathcal{N}(\underbrace{(\sigma_2^{-2}\boldsymbol{Z}_{1:t}\boldsymbol{Z}_{1:t}^T + \sigma_1^{-2}\boldsymbol{I}_N)^{-1}\sigma_2^{-2}\boldsymbol{Z}_{1:t}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\gamma})}_{\boldsymbol{\mu}_{**}}, (\sigma_2^{-2}\boldsymbol{Z}_{1:t}\boldsymbol{Z}_{1:t}^T + \sigma_1^{-2}\boldsymbol{I}_N)^{-1}).\end{aligned}$$

Using the Woodbury matrix identity, we have

$$\boldsymbol{\mu}_{**} = \sigma_1^2\boldsymbol{Z}_{1:t}(\sigma_2^2\boldsymbol{I}_{C_t} + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\gamma}).$$

Furthermore,

$$\begin{aligned}\sigma_2^{-2}\boldsymbol{Z}_{1:t}\boldsymbol{Z}_{1:t}^T + \sigma_1^{-2}\boldsymbol{I}_N &= \sigma_2^{-2}diag(n_t(1), \cdots, n_t(N)) + \sigma_1^{-2}\boldsymbol{I}_N\\ &= diag(\sigma_1^{-2} + \sigma_2^{-2}n_t(1), \cdots, \sigma_1^{-2} + \sigma_2^{-2}n_t(N)).\end{aligned}$$

Since $\boldsymbol{\theta} = \boldsymbol{\Phi}\boldsymbol{\gamma} + v$, then

$$\boldsymbol{\theta}|\boldsymbol{\gamma}, \mathcal{H}_{t+1} \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\gamma} + \boldsymbol{\mu}_{**}, diag(\sigma_1^{-2} + \sigma_2^{-2}n_t(1), \cdots, \sigma_1^{-2} + \sigma_2^{-2}n_t(N))^{-1}).$$

Therefore, for each item $i \in [N]$,

$$\begin{aligned}&\theta_i|\boldsymbol{\gamma}, \mathcal{H}_{t+1} \sim \mathcal{N}(\hat{\mu}_{t+1,\boldsymbol{\gamma}}(i), \hat{\sigma}_{t+1,\boldsymbol{\gamma}}^2(i)),\\ &\quad\hat{\mu}_{t+1,\boldsymbol{\gamma}}(i) = \mathbf{x}_i^T\boldsymbol{\gamma} + \sigma_1^2\boldsymbol{Z}_{1:t,i}(\sigma_2^2\boldsymbol{I}_{C_t} + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t})^{-1}(\boldsymbol{Y}_{1:t} - \boldsymbol{\Phi}_{1:t}\boldsymbol{\gamma}),\\ &\quad\hat{\sigma}_{t+1,\boldsymbol{\gamma}}^{-2}(i) = \sigma_1^{-2} + \sigma_2^{-2}n_t(i).\end{aligned} \tag{10}$$

C Preliminary and Definitions

We first clarify common notations used in our proof. Suppose that there are N items, each with d features. We will recommend a slate of at most K items each time. In total, there are T rounds of the interaction. Let us recall that $\mathcal{H}_t = (A_l, \boldsymbol{Y}_l(A_l))_{l=1}^{t-1}$ includes history up to round t and excluding round t, where $\mathcal{H}_1 = \emptyset$ and $\boldsymbol{Y}_l(A_l) = (Y_{k,l}, k \in A_l)$. Given the $\mathcal{H}_t$, the conditional probability is given as $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|\mathcal{H}_t)$, and the conditional expectation is given as $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot|\mathcal{H}_t)$. Similarly, we define the probability independent of all history as $\mathbb{P}(\cdot)$ and the expectation independent of all history as $\mathbb{E}(\cdot)$. Additionally, denote the number of pulls of arm k for the first t rounds (including round t) as $n_t(k)$. Suppose $\boldsymbol{X} \in \mathcal{R}^{d\times d}$, let $\lambda_1(\boldsymbol{X})$ denote the maximum eigenvalue of $\boldsymbol{X}$, and $\lambda_d(\boldsymbol{X})$ denote the minimum eigenvalue of $\boldsymbol{X}$.

We also need introduce some basic quantities from information theory. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures, and $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. Then the Kullback–Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined as $D(\mathbb{P}\|\mathbb{Q}) = \int log(\frac{d\mathbb{P}}{d\mathbb{Q}})d\mathbb{P}$, where $\frac{d\mathbb{P}}{d\mathbb{Q}}$ is the Radon–Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$. Then the mutual information between two random variables X and Y is defined as the Kullback–Leibler divergence between the joint distribution of X and Y and the product of the marginal distributions, $I(X;Y) = D(\mathbb{P}(X,Y)\|\mathbb{P}(X)\mathbb{P}(Y))$. The mutual information measures the information gained about one random variable by observing the other random variable, which is always non-negative and equals to 0 only if two random variables are independent to each other. For example, in the proof, we use $I(\boldsymbol{\gamma};\mathcal{H}_t)$ to quantify the information gain of $\boldsymbol{\gamma}$ by observing the historic interactions between agents and users, $\mathcal{H}_t$. We also need a conditional mutual information term to quantify the difference between random variables X and Y conditioned on another random variable Z, which is defined as $I(X;Y|Z) = \mathbb{E}[D(\mathbb{P}(X,Y|Z)\|\mathbb{P}(X|Z)\mathbb{P}(Y|Z))]$ (the expectation is taken over Z).

C.1 General History-Dependent Mutual Information

Conditional on history $\mathcal{H}_t$, the mutual information between the parameter $\boldsymbol{\theta}$ and the observations at round t, $\boldsymbol{Y}_t$, is defined as follows:

$$I_t(\boldsymbol{\theta}; A_t, \boldsymbol{Y}_t) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\boldsymbol{\theta}; A_t, \boldsymbol{Y}_t)}{\mathbb{P}_t(\boldsymbol{\theta})\mathbb{P}_t(A_t, \boldsymbol{Y}_t)}\Big)\Big].$$

Similarly, the history dependent mutual information between the meta parameter $\boldsymbol{\gamma}$ and the observations at round t, $\boldsymbol{Y}_t$, is defined as follows:

$$I_t(\boldsymbol{\gamma}; A_t, \boldsymbol{Y}_t) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\boldsymbol{\gamma}, A_t, \boldsymbol{Y}_t)}{\mathbb{P}_t(\boldsymbol{\gamma})\mathbb{P}_t(A_t, \boldsymbol{Y}_t)}\Big)\Big].$$

Then, the history dependent mutual information between the parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and the observations at round t, $\boldsymbol{Y}_t$, is defined as below:

$$I_t(\boldsymbol{\theta}, \boldsymbol{\gamma}; A_t, \boldsymbol{Y}_t) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\boldsymbol{\theta}, \boldsymbol{\gamma}, A_t, \boldsymbol{Y}_t)}{\mathbb{P}_t(\boldsymbol{\theta}, \boldsymbol{\gamma})\mathbb{P}_t(A_t, \boldsymbol{Y}_t)}\Big)\Big].$$

Finally, the history dependent mutual information between the parameters $\boldsymbol{\theta}$ and the observations at round t, $\boldsymbol{Y}_t$, given that the meta parameter $\boldsymbol{\gamma}$ is known, is defined as below:

$$I_t(\boldsymbol{\theta}; A_t, \boldsymbol{Y}_t|\boldsymbol{\gamma}) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\boldsymbol{\theta}, A_t, \boldsymbol{Y}_t|\boldsymbol{\gamma})}{\mathbb{P}_t(\boldsymbol{\theta}|\boldsymbol{\gamma})\mathbb{P}_t(A_t, \boldsymbol{Y}_t|\boldsymbol{\gamma})}\Big)\Big].$$

By the definition of conditional mutual information, we have

$$I(\cdot; A_t, \boldsymbol{Y}_t|\mathcal{H}_t) = \mathbb{E}(I_t(\cdot; A_t, \boldsymbol{Y}_t)),$$
$$I(\cdot; A_t, \boldsymbol{Y}_t|\boldsymbol{\gamma}, \mathcal{H}_t) = \mathbb{E}(I_t(\cdot; A_t, \boldsymbol{Y}_t|\boldsymbol{\gamma})).$$

C.2 History-Dependent/Independent Mutual Information and Entropy for Semi-Bandits

Conditional on history $\mathcal{H}_t$, the mutual information between the parameter θ_k and the observations at round t, $Y_{k,t}$, is defined as follows:

$$I_t(\theta_k; k, Y_{k,t}) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\theta_k; k, Y_{k,t})}{\mathbb{P}_t(\theta_k)\mathbb{P}_t(k, Y_{k,t})}\Big)\Big].$$

Similarly, the history dependent mutual information between the meta parameter γ and the observations at round t, $Y_{k,t}$, is defined as follows:

$$I_t(\gamma; k, Y_{k,t}) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\gamma, k, Y_{k,t})}{\mathbb{P}_t(\gamma)\mathbb{P}_t(k, Y_{k,t})}\Big)\Big].$$

Then, the history dependent mutual information between the parameters (θ_k, γ) and the observations at round t, $Y_{k,t}$, is defined as below:

$$I_t(\theta_k, \gamma; k, Y_{k,t}) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\theta_k, \gamma, k, Y_{k,t})}{\mathbb{P}_t(\theta_k, \gamma)\mathbb{P}_t(k, Y_{k,t})}\Big)\Big].$$

Finally, the history dependent mutual information between the parameters θ_k and the observations at round t, $Y_{k,t}$, given that the meta parameter γ is known, is defined as below:

$$I_t(\theta_k; k, Y_{k,t}|\gamma) = \mathbb{E}_t\Big[log\Big(\frac{\mathbb{P}_t(\theta_k, k, Y_{k,t}|\gamma)}{\mathbb{P}_t(\theta_k|\gamma)\mathbb{P}_t(k, Y_{k,t}|\gamma)}\Big)\Big].$$

Based on the definition of entropy, we further defined the history dependent entropy terms as follows:

$$\textbf{Conditional Entropy of } \theta_k : h_t(\theta_k) = -\mathbb{E}_t[log(\mathbb{P}_t(\theta_k))],$$
$$\textbf{Conditional Entropy of } \gamma : h_t(\gamma) = -\mathbb{E}_t[log(\mathbb{P}_t(\gamma))],$$
$$\textbf{Conditional Entropy of } \theta_k \textbf{ given } \gamma : h_t(\theta_k|\gamma) = -\mathbb{E}_t[log(\mathbb{P}_t(\theta_k|\gamma))].$$

Straightforwardly, by the definition of conditional mutual information, the history independent conditional mutual information terms are defined as the expectation of the history dependent term.

$$I(\cdot; k, Y_{k,t}|\mathcal{H}_t) = \mathbb{E}(I_t(\cdot; k, Y_{k,t})),$$
$$I(\cdot; k, Y_{k,t}|\gamma, \mathcal{H}_t) = \mathbb{E}(I_t(\cdot; k, Y_{k,t}|\gamma)).$$

Similarly, the history independent conditional entropy terms are defined as follows:

$$h(\cdot|\mathcal{H}_t) = \mathbb{E}(h_t(\cdot)),$$
$$h(\cdot|\gamma, \mathcal{H}_t) = \mathbb{E}(h_t(\cdot|\gamma)).$$

C.3 Others

In the following, we restate several properties of the mutual information and entropy and an inequality lemma that we mainly used in our proof.

Decomposition of Mutual Information. Based on the definition of mutual information and entropy, we can decompose the mutual information term as below.

$$I_t(\cdot; k, Y_{k,t}) = h_t(\cdot) - h_{t+1}(\cdot),$$
$$I_t(\cdot; k, Y_{k,t}|\gamma) = h_t(\cdot|\gamma) - h_{t+1}(\cdot|\gamma).$$

Chain Rule. $I(X, Y; Z) = I(Y; Z) + I(X; Z|Y)$.

Weyl's inequality. For Hermitian matrix $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{d\times d}$ and $i = 1, \cdots, d$,

$$\lambda_i(\boldsymbol{A} + \boldsymbol{B}) \leq \lambda_i(\boldsymbol{A}) + \lambda_1(\boldsymbol{B}).$$

D Main Proof

D.1 Proof for Theorem 1

Proof. First, following the property of mutual information and the chain rule of conditional mutual information, we can derive that $I_t(\boldsymbol{\theta}; A_t, \boldsymbol{Y}_t) \leq I_t(\boldsymbol{\theta}, \gamma; A_t, \boldsymbol{Y}_t) = I_t(\gamma; A_t, \boldsymbol{Y}_t) + I_t(\boldsymbol{\theta}; A_t, \boldsymbol{Y}_t|\gamma)$. Taking the square root of it and applying the Cauchy-Schwartz inequality, we have that $\sqrt{I_t(\boldsymbol{\gamma}; A_t, \mathbf{Y}_t) + I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t|\boldsymbol{\gamma})} \leq \sqrt{I_t(\boldsymbol{\gamma}; A_t, \mathbf{Y}_t)} + \sqrt{I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t|\boldsymbol{\gamma})}$. After that, using the assumption that $\Gamma_t \leq \Gamma\ w.p.1$ and collecting the terms, we finish the proof. Here, the regret bound can be divided into two parts, where the first part is the cost of learning the meta parameter $\boldsymbol{\gamma}$, the second part is the regret for learning $\boldsymbol{\theta}$ with known $\boldsymbol{\gamma}$.

Mathematically,

$$\begin{aligned}
BR(T) &= \mathbb{E}[\sum_t \Delta_t] \\
&\leq \mathbb{E}[\sum_t \Gamma_t \sqrt{I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t)} + \epsilon_t] \\
&\leq \mathbb{E}[\sum_t \Gamma_t \sqrt{I_t(\boldsymbol{\theta}, \boldsymbol{\gamma}; A_t, \mathbf{Y}_t)}] + \mathbb{E}[\sum_t \epsilon_t] \\
&= \mathbb{E}[\sum_t \Gamma_t \sqrt{I_t(\boldsymbol{\gamma}; A_t, \mathbf{Y}_t) + I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t|\boldsymbol{\gamma})}] + \mathbb{E}[\sum_t \epsilon_t] \\
&\leq \mathbb{E}[\Gamma_t \sum_t \sqrt{I_t(\boldsymbol{\gamma}; A_t, \mathbf{Y}_t)} + \sqrt{I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t|\boldsymbol{\gamma})}] + \mathbb{E}[\sum_t \epsilon_t] \\
&\leq \underbrace{\Gamma \sum_t \mathbb{E}[\sqrt{I_t(\boldsymbol{\gamma}; A_t, \mathbf{Y}_t)}]}_{\text{Regret due to not knowing } \boldsymbol{\gamma}} + \underbrace{\sum_t \Gamma \mathbb{E}[\sqrt{I_t(\boldsymbol{\theta}; A_t, \mathbf{Y}_t|\boldsymbol{\gamma})}] + \mathbb{E}[\epsilon_t]}_{\text{Regret suffered even with known } \gamma}.
\end{aligned}$$

The first inequality directly uses the (4). The second inequality follows the property of mutual information that $I(X; Z) \leq I(X, Y; Z)$. Here $X = \boldsymbol{\theta}$, $Y = \boldsymbol{\gamma}$, and $Z = (A_t, \mathbf{Y}_t)$. The third equality uses the chain rule of mutual information, $I(X, Y; Z) = I(Y; Z) + I(X; Z|Y)$. The forth inequality follows the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. The final inequality follows that $\Gamma_t \leq \Gamma\ w.p.1$. □

D.2 Proof for Theorem 2

Roadmap: *There are two main steps in the proof. First, we decompose the Bayes regret into two parts as (12). To derive the Bayes regret decomposition, we first show that (11) holds for all $t \in [T]$ in Lemma 1, and then prove that (12) holds under the condition (11) in Lemma 2. Second, we get the bound of each component in (12). In particular, the upper bounds of Γ_t and ϵ_t are derived in Lemma 1, whereas the upper bounds of $I(\gamma; \mathcal{H}_{T+1})$ and $I(\theta_k; \mathcal{H}_{T+1})$ are derived in Lemma 3. Gathering the bounds of all components, we get the regret bound in Theorem 2. Following are the details of the main proof.*

We start by stating several lemmas, which will be used in our main proof. Proofs of the lemmas are deferred to Appendix E. Without loss of generality, we assume that all available items have bounded norm (Assumption 1) and all parameters are bounded (Assumption 2).

Using the independence between rewards generated by different arms, we first decompose the per-round expected regret in a similar form of (4), with suitably selected history-dependent constants Γ_t and ϵ_t. Based on the properties of the Gaussian distributions and the fact that MTSS samples rewards from corresponding posterior distributions for every round, we bound both Γ_t and ϵ_t by functions of $\frac{\delta}{N} \in (0, 1]$.

Lemma 1. *For any $\mathcal{H}_t$-adapted sequence of actions $(A_l)_{l=1}^{t-1}$, and any δ such that $\frac{\delta}{N} \in (0, 1]$, the expected regret in round t conditioned on $\mathcal{H}_t$ is bounded as*

$$\mathbb{E}_t[\Delta_t] \leq \sum_{k \in [N]} \mathbb{P}_t(k \in A_t) \Gamma_{k,t} \sqrt{I_t(\theta_k; k, Y_{k,t})} + \epsilon_t, \tag{11}$$

where

$$\Gamma_{k,t} = 4\sqrt{\frac{\hat{\sigma}_t^2(k)}{log(1 + \hat{\sigma}_t^2(k)/\sigma_2^2)} log(\frac{4N}{\delta})}, \qquad \epsilon_t = \sum_{k \in [N]} \mathbb{P}_t(k \in A_t) \sqrt{2\delta \frac{1}{N} \hat{\sigma}_t^2(k)}.$$

Moreover, for each k, the following history-independent bound holds almost surely.

$$\hat{\sigma}_t^2(k) \leq \lambda_1(\Sigma_{\boldsymbol{\gamma}}) + \sigma_1^2.$$

Based on **Lemma 1**, we get that $\Gamma_{k,t} = O(\sqrt{log(\frac{N}{\delta})})$ and $\epsilon_t = O(K\sqrt{\frac{\delta}{N}})$. Then, similar to **Theorem 1**, based on the per-round conditional expected regret decomposition, we develop a decomposition of the total regret over T rounds of interactions by summing the per-round regret over T rounds and then taking the expectation over historical interactions.

Lemma 2. *Suppose that (11) holds for all $t \in [T]$, for some suitably chosen $\Gamma_{k,t}$ and ϵ_t. Let Γ_k and Γ be some non-negative constants such that $\Gamma_{k,t} \leq \Gamma_k \leq \Gamma$ holds for all $t \in [T]$ and $k \in [N]$ almost surely. Then*

$$BR(T) \leq \underbrace{\Gamma K\sqrt{TI(\gamma; \mathcal{H}_{T+1})}}_{\textit{Regret due to not knowing } \gamma} + \underbrace{\Gamma\sqrt{NTK}\sqrt{\frac{1}{N}\sum_{k\in[N]} I(\theta_k; \mathcal{H}_{T+1}|\gamma)} + \sum_t \mathbb{E}[\epsilon_t]}_{\textit{Regret suffered even with known } \gamma}. \tag{12}$$

Here, the first term is the cost for learning the meta parameter $\boldsymbol{\gamma}$, and the second term is regret for unknown item-specific parameter θ_k given known $\boldsymbol{\gamma}$. We show the benefits of information sharing among items mainly by the first term, which indicates that the extra regret due to unknown $\boldsymbol{\gamma}$ is much lower that the cost of learning $\boldsymbol{\theta}$ with known $\boldsymbol{\gamma}$. Using the assumption that $\Gamma_{k,t} \leq \Gamma_k \leq \Gamma \ w.p.1$ and the bound of $\Gamma_{k,t}$ and ϵ_t, we directly get the bound of Γ and $\mathbb{E}[\epsilon_t]$. Then, our next lemma find the bound of the mutual information terms involved in (12), by using the properties of Gaussian distribution and the properties of LMM.

Lemma 3. *For any $k \in [N]$ and any $\mathcal{H}_{T+1}$-adapted sequence of actions $(A_l)_{l=1}^T$, we have*

$$I(\gamma; \mathcal{H}_{T+1}) \leq \frac{d}{2}log\Big(1 + \frac{N\lambda_1(\Sigma_\gamma)}{\sigma_1^2 + \sigma_2^2/T}\Big), \qquad I(\theta_k; \mathcal{H}_{T+1}|\gamma) \leq \frac{1}{2}log(1 + \frac{\sigma_1^2}{\sigma_2^2}T).$$

Now we are ready to combine these results and present our main proof of **Theorem 2**. Specifically, we get the bounds of $\Gamma_{k,t}$ and ϵ_t from **Lemma 1** and the bounds of mutual information terms from **Lemma 3**, and then plug them into the regret decomposition derived in **Lemma 2**.

Proof of Theorem 2. From **Lemma 1**, we showed that (11) holds for suitably chosen $\Gamma_{k,t}$ and ϵ_t. Using the upper bounds of $\hat{\sigma}_t(k)$ in **Lemma 1**, since $\sqrt{\frac{x}{log(1+ax)}}$ is an increasing function in x when $a > 0$, we can bound $w.p.1$ that

$$\Gamma_{k,t} \leq 4\sqrt{\frac{\lambda_1(\Sigma_\gamma) + \sigma_1^2}{log(1 + (\lambda_1(\Sigma_\gamma) + \sigma_1^2)/\sigma_2^2)}log(\frac{4N}{\delta})} = \Gamma.$$

Then, we have the upper bound of $\Gamma_{k,t} \leq \Gamma$ for all t and k $w.p.1$. Similarly, we have

$$\epsilon_t \leq \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\sqrt{2\delta\frac{1}{N}(\lambda_1(\Sigma_\gamma) + \sigma_1^2)}.$$

From **Lemma 2**, for any $\delta > 0$, let $c_1 = 4\sqrt{\frac{\lambda_1(\Sigma_\gamma)+\sigma_1^2}{log(1+(\lambda_1(\Sigma_\gamma)+\sigma_1^2)/\sigma_2^2)}log(\frac{4N}{\delta})}$, we have

$$\begin{aligned} BR(T) \leq & \Gamma K\sqrt{TI(\gamma; \mathcal{H}_{T+1})} + \Gamma\sqrt{NTK}\sqrt{\frac{1}{N}\sum_{k\in[N]} I(\theta_k; \mathcal{H}_{T+1}|\gamma) + \sum_t \mathbb{E}(\epsilon_t)} \\ \leq & c_1 K\sqrt{T}\sqrt{\frac{d}{2}log\Big(1 + \frac{N\lambda_1(\Sigma_\gamma)}{\sigma_1^2 + \sigma_2^2/T}\Big)} + c_1\sqrt{NTK}\sqrt{\frac{1}{2}log(1 + \frac{\sigma_1^2}{\sigma_2^2}T)} \\ & + TK\sqrt{2\delta\frac{1}{N}(\lambda_1(\Sigma_\gamma) + \sigma_1^2)} \end{aligned}$$

The inequality holds by first using the upper bound of mutual information in **Lemma 3**, and the upper bound of Γ and ϵ_t, then we derive the history-independent upper bound of $\sum_t \mathbb{E}(\epsilon_t)$ as the following.

$$\begin{aligned}
&\sum_t \mathbb{E}\Big[\sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\sqrt{2\delta\frac{1}{N}(\lambda_1(\Sigma_\gamma)+\sigma_1^2)}\Big] \\
&= \sqrt{2\delta\frac{1}{N}(\lambda_1(\Sigma_\gamma)+\sigma_1^2)}\mathbb{E}\Big[\sum_t\sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\Big] \\
&\leq TK\sqrt{2\delta\frac{1}{N}(\lambda_1(\Sigma_\gamma)+\sigma_1^2)}.
\end{aligned}$$

Let $\delta = \frac{1}{T^2}$,

$$\begin{aligned}
BR(T) \leq & c_1 K\sqrt{T}\sqrt{\frac{d}{2}log\Big(1+\frac{N\lambda_1(\Sigma_\gamma)}{\sigma_1^2+\sigma_2^2/T}\Big)} + c_1\sqrt{NTK}\sqrt{\frac{1}{2}log(1+\frac{\sigma_1^2}{\sigma_2^2}T)} \\
&+ K\sqrt{\frac{2}{N}(\lambda_1(\Sigma_\gamma)+\sigma_1^2)} \\
=& O(K\sqrt{Tdlog(N)log(NT^2)} + \sqrt{NTKlog(T)log(NT^2)} + K\sqrt{\frac{1}{N}}) \\
=& \tilde{O}(K\sqrt{Td} + \sqrt{NTK}).
\end{aligned}$$

□

E Proof of Lemmas

E.1 Proof for Lemma 1

Proof. First, using the probability matching property of Thompson Sampling and the independence between the rewards generated by different arms, we decompose the per-round expected regret as $\sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\mathbb{E}_t[\hat{\theta}_{k,t} - \theta_k]$, where $\hat{\theta}_{k,t}$ is the estimated mean reward for arm k given the history $\mathcal{H}_t$. Then, following Lemma 5 in Lu and Van Roy [2019], we define a confidence set $\Theta_t(k)$ for both $\hat{\theta}_{k,t}$ and θ_k with high probability for each arm k at round t, with suitably selected non-negative random variables $\Gamma_{k,t}$, which leads to the bound of $\mathbb{E}_t[\hat{\theta}_{k,t} - \theta_k]$ and concludes the proof of the first part of the lemma directly. The ϵ_t is some non-negative random variables derived appropriately. For the second part of the lemma, we bound the $\Gamma_{k,t}$ and ϵ_t by finding the upper bound of $\hat{\sigma}_t^2(k)$ for each arm k at round t conditional on the history $\mathcal{H}_t$.

Now we are ready to prove **Lemma 1** in detail, as follows.

Since $\sum_{k\in A_*} \theta_k|\mathcal{H}_t \overset{d}{=} \sum_{k\in A_t} \hat{\theta}_{k,t}|\mathcal{H}_t$, we have

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] &= \mathbb{E}_t[\sum_{k\in A_*}\theta_k - \sum_{k\in A_t}\theta_k] \\
&= \mathbb{E}_t[\sum_{k\in A_t}\hat{\theta}_{k,t} - \sum_{k\in A_t}\theta_k] \\
&= \mathbb{E}_t[\sum_{k\in[N]}\mathbb{1}(k\in A_t)(\hat{\theta}_{k,t} - \theta_k)] \\
&= \sum_{k\in[N]}\mathbb{P}_t(k\in A_t)\mathbb{E}_t[\hat{\theta}_{k,t} - \theta_k].
\end{aligned}$$

For each $k \in [N]$, we know that $\hat{\theta}_{k,t}|\mathcal{H}_t \sim \mathcal{N}(\hat{\mu}_t(k), \hat{\sigma}_t^2(k))$. Let us consider the confidence set of $\hat{\theta}_k$ for each arm k:

$$\Theta_t(k) = \{\theta : |\theta - \hat{\mu}_t(k)| \leq \frac{\Gamma_{k,t}}{2}\sqrt{I_t(\theta_k; k, Y_{k,t})}\}.$$

The history dependent conditional mutual entropy of θ_k given the history $\mathcal{H}_t$, $I_t(\theta_k; k, Y_{k,t})$, can be computed as follows:

$$\begin{aligned}
I_t(\theta_k; k, Y_{k,t}) &= h_t(\theta_k) - h_{t+1}(\theta_k) \\
&= \frac{1}{2} log(det(2\pi e \hat{\sigma}_t^2(k)) - \frac{1}{2} log(det(2\pi e \hat{\sigma}_{t+1}^2(k)) \\
&= \frac{1}{2} log(\hat{\sigma}_t^2(k) \hat{\sigma}_{t+1}^{-2}(k)) \\
&= \frac{1}{2} log(\hat{\sigma}_t^2(k)[\hat{\sigma}_t^{-2}(k) + \sigma_2^{-2}]) \\
&= \frac{1}{2} log\Big(1 + \frac{\hat{\sigma}_t^2(k)}{\sigma_2^2}\Big).
\end{aligned}$$

For $\frac{\delta}{N} > 0$, let

$$\Gamma_{k,t} = 4\sqrt{\frac{\hat{\sigma}_t^2(k)}{log(1 + \hat{\sigma}_t^2(k)/\sigma_2^2)} log(\frac{4N}{\delta})}.$$

Then, following Lemma 5 in Lu and Van Roy [2019], for any k and any δ such that $\frac{\delta}{N} \in (0, 1]$, we have

$$\mathbb{P}_t(\hat{\theta}_{k,t} \in \Theta_t(k)) \geq 1 - \frac{\delta}{2N}.$$

Now we continue the regret decomposition as

$$\begin{aligned}
\mathbb{E}_t[\hat{\theta}_{k,t} - \theta_k] &= \mathbb{E}_t[\mathbb{1}(\hat{\theta}_{k,t}, \theta_k \in \Theta_t(k))(\hat{\theta}_{k,t} - \theta_k)] + \mathbb{E}_t[\mathbb{1}^c(\hat{\theta}_{k,t}, \theta_k \in \Theta_t(k))(\hat{\theta}_{k,t} - \theta_k)] \\
&\leq \Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})} + \sqrt{\mathbb{P}(\hat{\theta}_{k,t} \text{ or } \theta_k \notin \Theta_t(k))\mathbb{E}_t[(\hat{\theta}_{k,t} - \theta_k)^2]} \\
&\leq \Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})} + \sqrt{\delta\frac{1}{N}\mathbb{E}_t[(\hat{\theta}_{k,t} - \hat{\mu}_t(k))^2 + (\theta_k - \hat{\mu}_t(k))^2]} \\
&\leq \Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})} + \sqrt{2\delta\frac{1}{N}\hat{\sigma}_t^2(k)}.
\end{aligned}$$

The second inequality uses that $\mathbb{P}(\hat{\theta}_{k,t} \text{ or } \theta_k \notin \Theta_t(k)) \leq \mathbb{P}(\hat{\theta}_{k,t} \notin \Theta_t(k)) + \mathbb{P}(\theta_k \notin \Theta_t(k)) = \frac{\delta}{N}$. Therefore, we conclude our proof of the first part.

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] &= \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\mathbb{E}_t[\hat{\theta}_{k,t} - \theta_k] \\
&\leq \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\Big(\Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})} + \sqrt{2\delta\frac{1}{N}\hat{\sigma}_t^2(k)}\Big) \\
&= \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})} + \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\sqrt{2\delta\frac{1}{N}\hat{\sigma}_t^2(k)} \\
&= \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})} + \epsilon_t.
\end{aligned}$$

Bounding $\hat{\sigma}_t^2(k)$. Recall that

$$\begin{aligned}
\hat{\sigma}_t^2(k) &= \mathbf{x}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\mathbf{x}_i + \sigma_1^2 - (\mathbf{x}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t,i}) \\
&\quad\times(\sigma_2^2\boldsymbol{I}_{Kt} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t}^T\boldsymbol{Z}_{1:t})^{-1}(\mathbf{x}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t,i})^T \\
&\leq \mathbf{x}_k^T\Sigma_{\boldsymbol{\gamma}}\mathbf{x}_k + \sigma_1^2 \\
&\leq max_{k\in[N]}\mathbf{x}_k^T\Sigma_{\boldsymbol{\gamma}}\mathbf{x}_k + \sigma_1^2 \\
&\leq max_{k\in[N]}\mathbf{x}_k^T\lambda_1(\Sigma_{\boldsymbol{\gamma}})\mathbf{x}_k + \sigma_1^2 \\
&\leq \lambda_1(\Sigma_{\boldsymbol{\gamma}}) + \sigma_1^2.
\end{aligned}$$

The first inequality results form that

$$(\mathbf{x}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{\Phi}_{1:t}^T + \sigma_1^2 \boldsymbol{Z}_{1:t,i})(\sigma_2^2 \boldsymbol{I}_{Kt} + \boldsymbol{\Phi}_{1:t}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2 \boldsymbol{Z}_{1:t}^T \boldsymbol{Z}_{1:t})^{-1}(\mathbf{x}_i^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{\Phi}_{1:t}^T + \sigma_1^2\boldsymbol{Z}_{1:t,i})^T$$

is positive semi-definite. By the assumption that $\|x_i\|_2 \leq 1$, we have the last inequality. □

E.2 Proof for Lemma 2

Proof. First, suppose that (11) holds for all round t, we adapt **Theorem 1** to bound the expected regret for each arm k. Here, $A_t = k$, $\boldsymbol{Y}_t = Y_{k,t}$, and $\Gamma_t = \mathbb{P}_t(k \in A_t)\Gamma_{k,t}$. Summing the regret bound for each arm k, similar to **Theorem 1**, we can decompose the Bayes regret bound into three parts, where the first part is the cost of learning $\boldsymbol{\gamma}$, and the rest two parts constitute the cost of learning $\boldsymbol{\theta} = (\theta_k)_{k=1}^N$. While the third part is bounded in **Lemma 1**, we bound the first two parts separately. Particularly, by **Lemma 4**, the expectation of the history dependent mutual information terms in the first two parts are bounded by the history independent mutual information, respectively. The proof is concluded by utilizing inequalities and the assumption that $\Gamma_{k,t} \leq \Gamma_k \leq \Gamma\ w.p.1$.

Mathematically,

$$\begin{aligned}
BR(T) =& \mathbb{E}[R(T)] = \mathbb{E}[\sum_t \Delta_t] \\
\leq& \mathbb{E}\Big[\sum_t \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})}\Big] + \mathbb{E}\sum_t \epsilon_t \\
=& \sum_{k\in[N]} \mathbb{E}\Big[\sum_t \mathbb{P}_t(k \in A_t)\Gamma_{k,t}\sqrt{I_t(\theta_k; k, Y_{k,t})}\Big] + \mathbb{E}\sum_t \epsilon_t \\
\leq& \underbrace{\sum_{k\in[N]} \Gamma_k \sum_t \mathbb{E}[\mathbb{P}_t(k \in A_t)\sqrt{I_t(\boldsymbol{\gamma}; k, Y_{k,t})}]}_{\text{Regret due to not knowing } \boldsymbol{\gamma}} \\
&+ \underbrace{\sum_{k\in[N]} \Gamma_k \sum_t \mathbb{E}[\mathbb{P}_t(k \in A_t)\sqrt{I_t(\theta_k; k, Y_{k,t}|\boldsymbol{\gamma})}] + \sum_t \mathbb{E}[\epsilon_t]}_{\text{Regret suffered even with known } \gamma}.
\end{aligned}$$

The first inequality directly uses the (11). Similar to the proof of **Theorem 1**, we have the second inequality with the assumption that $\Gamma_{k,t} \leq \Gamma_k\ w.p.1$.

We now derive the bounds of the first two terms separately. For the first term, we have

$$\begin{aligned}
&\sum_{k\in[N]} \Gamma_k \mathbb{E}\sum_t \mathbb{P}_t(k \in A_t)\sqrt{I_t(\boldsymbol{\gamma}; k, Y_{k,t})} \\
&\leq \Gamma\mathbb{E}\sum_t \sum_{k\in[N]} \mathbb{P}_t(k \in A_t)\sqrt{I_t(\boldsymbol{\gamma}; k, Y_{k,t})} \\
&\leq \Gamma K \mathbb{E}\sum_t \sqrt{I_t(\boldsymbol{\gamma}; A_t, \boldsymbol{Y}_t)} \\
&\leq \Gamma K E\sqrt{T\sum_t I_t(\boldsymbol{\gamma}; A_t, \boldsymbol{Y}_t)} \\
&\leq \Gamma K\sqrt{T\sum_t \mathbb{E}I_t(\boldsymbol{\gamma}; A_t, \boldsymbol{Y}_t)} \\
&= \Gamma K\sqrt{TI(\boldsymbol{\gamma}; \mathcal{H}_{T+1})}.
\end{aligned}$$

The first inequality follows that $\Gamma_k \leq \Gamma\ w.p.1$. The second inequality holds by **Lemma 4**. The third inequality follows $\sum_i a_i b_i \leq \sqrt{\sum_i a_i^2 \sum_i b_i^2}$ with $a_i = 1$ and $b_i = \sqrt{I_t(\boldsymbol{\gamma}; A_t, Y_t)}$. The forth inequality follows the Jensen's Inequality for $\sqrt{\cdot}$. The final equality follows the chain rule of mutual information. Specifically, $I(\boldsymbol{\gamma}; \mathcal{H}_{T+1}) = \sum_{t=1}^T I(\boldsymbol{\gamma}; A_t, \boldsymbol{Y}_t|\mathcal{H}_t) = \mathbb{E}\sum_{t=1}^T I_t(\boldsymbol{\gamma}; A_t, \boldsymbol{Y}_t)$.

For the second term, we have

$$\begin{aligned}
&\sum_{k\in[N]}\Gamma_k\mathbb{E}\sum_t\mathbb{P}_t(k\in A_t)\sqrt{I_t(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma})}\\
&=\sum_{k\in[N]}\Gamma_k\mathbb{E}\Big[\sum_t\sqrt{\mathbb{P}_t(k\in A_t)}\sqrt{\mathbb{P}_t(k\in A_t)I_t(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma})}\Big]\\
&\le\sum_{k\in[N]}\Gamma_k\mathbb{E}\Big[\sqrt{\sum_t\mathbb{P}_t(k\in A_t)\sum_t\mathbb{P}_t(k\in A_t)I_t(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma})}\Big]\\
&\le\sum_{k\in[N]}\Gamma_k\sqrt{\mathbb{E}\sum_t\mathbb{P}_t(k\in A_t)}\sqrt{\mathbb{E}\sum_t\mathbb{P}_t(k\in A_t)I_t(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma})}\\
&\le\sum_{k\in[N]}\Gamma_k\sqrt{\mathbb{E}[n_T(k)]}\sqrt{I(\theta_k;\mathcal{H}_{T+1}|\boldsymbol{\gamma})}\\
&\le\sqrt{N\sum_{k\in[N]}\mathbb{E}[n_T(k)]}\sqrt{\frac{1}{N}\sum_{k\in[N]}\Gamma_k^2I(\theta_k;\mathcal{H}_{T+1}|\boldsymbol{\gamma})}\\
&\le\sqrt{NTK}\sqrt{\frac{1}{N}\sum_{k\in[N]}\Gamma^2I(\theta_k;\mathcal{H}_{T+1}|\boldsymbol{\gamma})}\\
&=\Gamma\sqrt{NTK}\sqrt{\frac{1}{N}\sum_{k\in[N]}I(\theta_k;\mathcal{H}_{T+1}|\boldsymbol{\gamma})}.
\end{aligned}$$

The first inequality uses the Cauchy-Schwartz inequality, that $\sum_i a_ib_i\le\sqrt{\sum_i a_i^2\sum_i b_i^2}$. Here $a_i=\sqrt{\mathbb{P}_t(k\in A_t)}$ and $b_i=\sqrt{\mathbb{P}_t(k\in A_t)I_t(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma})}$. The second inequality follows that $\mathbb{E}(XY)\le\sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$ for $X,Y>0\ w.p.1$. The third inequality uses the result of **Lemma 4**. The next inequality uses the Cauchy-Schwartz inequality again with $a_i=\sqrt{\mathbb{E}[n_T(k)]}$ and $b_i=\sqrt{I(\theta_k;\mathcal{H}_{T+1}|\boldsymbol{\gamma})}$. The last inequality is because of $\Gamma_k\le\Gamma\ w.p.1$. □

E.3 Proof for Lemma 3

Proof. First, we derive the mutual information of the meta-parameter $\boldsymbol{\gamma}$ given the history as follows.

$$\begin{aligned}
I(\boldsymbol{\gamma};\mathcal{H}_{T+1})&=h(\boldsymbol{\gamma})-h(\boldsymbol{\gamma}|\mathcal{H}_{T+1})\\
&=h(\boldsymbol{\gamma})-\mathbb{E}[h_{T+1}(\boldsymbol{\gamma})]\\
&=\frac{1}{2}log(det(2\pi e\Sigma_{\boldsymbol{\gamma}}))-\mathbb{E}[\frac{1}{2}log(det(2\pi e\tilde{\Sigma}_{T+1}))]\\
&=\frac{1}{2}\mathbb{E}[log(det(\Sigma_{\boldsymbol{\gamma}})det(\tilde{\Sigma}_{T+1}^{-1}))]\\
&=\frac{1}{2}\mathbb{E}[log(\prod_{i=1}^{d}\lambda_i(\Sigma_{\boldsymbol{\gamma}})\lambda_i(\tilde{\Sigma}_{T+1}^{-1}))]\\
&\le\frac{1}{2}\mathbb{E}\Big[log(\prod_{i=1}^{d}\lambda_i(\Sigma_{\boldsymbol{\gamma}})\Big(\frac{1}{\lambda_i(\Sigma_{\boldsymbol{\gamma}})}+\frac{N}{\sigma_2^2/T+\sigma_1^2}\Big)\Big]\\
&\le\frac{d}{2}log\Big(1+\frac{N\lambda_1(\Sigma_{\boldsymbol{\gamma}})}{\sigma_1^2+\sigma_2^2/T}\Big).
\end{aligned}$$

For the final inequality, we derive the history independent bound as follows.

$$\begin{aligned}\lambda_i(\tilde{\mathbf{\Sigma}}_{T+1}^{-1})) &= \lambda_i(\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{-1} + \sum_{i=1}^{N} \frac{n_T(i)}{\sigma_2^2 + \sigma_1^2 n_T(i)} \mathbf{x}_i \mathbf{x}_i^T) \\ &\le \lambda_i(\mathbf{\Sigma}_{\boldsymbol{\gamma}}^{-1}) + \lambda_1(\sum_{i=1}^{N} \frac{n_T(i)}{\sigma_2^2 + \sigma_1^2 n_T(i)} \mathbf{x}_i \mathbf{x}_i^T) \\ &\le \frac{1}{\lambda_i(\mathbf{\Sigma}_{\boldsymbol{\gamma}})} + tr(\sum_{i=1}^{N} \frac{n_T(i)}{\sigma_2^2 + \sigma_1^2 n_T(i)} \mathbf{x}_i \mathbf{x}_i^T) \\ &= \frac{1}{\lambda_i(\mathbf{\Sigma}_{\boldsymbol{\gamma}})} + \sum_{i=1}^{N} \frac{n_T(i)}{\sigma_2^2 + \sigma_1^2 n_T(i)} tr(\mathbf{x}_i^T \mathbf{x}_i) \\ &\le \frac{1}{\lambda_i(\mathbf{\Sigma}_{\boldsymbol{\gamma}})} + \sum_{i=1}^{N} \frac{n_T(i)}{\sigma_2^2 + \sigma_1^2 n_T(i)} \\ &\le \frac{1}{\lambda_i(\mathbf{\Sigma}_{\boldsymbol{\gamma}})} + \frac{N}{\sigma_2^2/T + \sigma_1^2}.\end{aligned}$$

The first inequality follows the Weyl's inequality. The second equality first uses linearity of trace, and then uses the cyclic property of trace. By assumption 1, we have $tr(\mathbf{x}_i^T \mathbf{x}_i) = \|\mathbf{x}_i\|_2^2 \le 1$, and the second last inequality holds.

Now we derive the mutual information of θ_k for each item $k \in [N]$, given the history and the meta-parameter $\boldsymbol{\gamma}$.

$$\begin{aligned}I(\theta_k; \mathcal{H}_{T+1}|\boldsymbol{\gamma}) &= h(\theta_k|\boldsymbol{\gamma}) - h(\theta_k|\boldsymbol{\gamma}, \mathcal{H}_{T+1}) \\ &= h(\theta_k|\boldsymbol{\gamma}) - \mathbb{E}[h_{T+1}(\theta_k|\boldsymbol{\gamma})] \\ &= \frac{1}{2} log(det(2\pi e \sigma_1^2)) - \mathbb{E}[\frac{1}{2} log(det(2\pi e \hat{\sigma}_{T+1,\gamma}^2(k))] \\ &= \frac{1}{2}\mathbb{E}\Big[log[\sigma_1^2(\sigma_1^{-2} + \sigma_2^{-2} n_T(k))]\Big] \\ &= \frac{1}{2}\mathbb{P}(n_T(k) \ge 1)\mathbb{E}\Big[log[1 + \frac{\sigma_1^2}{\sigma_2^2} n_T(k)]\Big] \\ &\le \frac{1}{2} log(1 + \frac{\sigma_1^2}{\sigma_2^2}\mathbb{E}[n_T(k)]) \\ &\le \frac{1}{2} log(1 + \frac{\sigma_1^2}{\sigma_2^2} T).\end{aligned}$$

The first inequality first uses the fact that $\mathbb{P}(n_T(k) \ge 1) \le 1$, then follows the Jensen's inequality of log. □

E.4 Proof for Lemma 4

Lemma 4. *For any $k \in [N]$ and $\mathcal{H}_t$-adapted sequence of actions $(A_l)_{l=1}^{t-1}$, the following statements hold*

$$\begin{aligned}I(\theta_k; \mathcal{H}_{T+1}|\boldsymbol{\gamma}) &= \mathbb{E}\sum_{t} \mathbb{P}_t(k \in A_t) I_t(\theta_k; k, Y_{k,t}|\boldsymbol{\gamma}), \\ K\sqrt{I_t(\boldsymbol{\gamma}; A_t, \mathbf{Y}_t)} &\ge \sum_{k \in [N]} \mathbb{P}_t(k \in A_t)\sqrt{I_t(\boldsymbol{\gamma}; k, Y_{k,t})}.\end{aligned}$$

Proof. First, we derive the conditional mutual information of θ_k given history and the meta-parameter $\boldsymbol{\gamma}$. Note that, in the rounds when arm k was not played, the mutual information gain for θ_k given $\boldsymbol{\gamma}$ is *zero*. In order words, $I_t(\theta_k; k, Y_{k,t}|\boldsymbol{\gamma}) = 0$ if arm k was not played at round t. Then we used the chain rule of the mutual information ($I(X; Y, Z) = I(X; Z) +$

$I(X;Y|Z)$) to finish the proof.

$$\begin{aligned}
I(\theta_k;\mathcal{H}_{T+1}|\boldsymbol{\gamma}) &= \sum_t I(\theta_k;A_t,\boldsymbol{Y}_t|\boldsymbol{\gamma},\mathcal{H}_t)\\
&= \mathbb{E}\sum_t I_t(\theta_k;A_t,\boldsymbol{Y}_t|\boldsymbol{\gamma})\\
&= \mathbb{E}\sum_t\sum_{a\in\mathcal{A}}\mathbb{P}_t(A_t=a)I_t(\theta_k;a,\boldsymbol{Y}_t(a)|\boldsymbol{\gamma})\\
&= \mathbb{E}\sum_t\sum_{a\in\mathcal{A}}\mathbb{P}_t(A_t=a)\mathbb{1}(k\in a)I_t(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma})\\
&+ \mathbb{E}\sum_t\sum_{a\in\mathcal{A}}\mathbb{P}_t(A_t=a)I_t(\theta_k;a\neg k,Y_t(a\neg k)|\boldsymbol{\gamma},(k,Y_{k,t}))\\
&= \mathbb{E}\sum_t\mathbb{P}_t(k\in A_t)I_t(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma}).
\end{aligned}$$

$a\neg k$ indicates that arm k is removed from action set a, and $\boldsymbol{Y}_t(a)$ is the observed rewards of set a. The last inequality follows that, given $(\theta_k;k,Y_{k,t}|\boldsymbol{\gamma})$, history and $\boldsymbol{\gamma}$, $\theta_k \perp\!\!\!\perp (a\neg k,Y_t(a\neg k))$, and $\mathbb{E}\sum_t\sum_{a\in\mathcal{A}}\mathbb{P}_t(A_t = a)I_t(\theta_k;a\neg k,Y_t(a\neg k)|\boldsymbol{\gamma},(k,Y_{k,t}))=0$.

For the second part of the lemma, we use the fact that $I(X;Y,Z)\geq max(I(X;Z),I(X,Y))$, which is intuitive, as more observations will provide more mutual information gain. For a fixed A_t, we have

$$\begin{aligned}
\sum_{k\in[N]}\mathbb{P}_t(k\in A_t)\sqrt{I_t(\boldsymbol{\gamma};k,Y_{k,t})} &\leq \sum_{k\in[N]}\mathbb{P}_t(k\in A_t)\sqrt{I_t(\boldsymbol{\gamma};A_t,\boldsymbol{Y}_t)}\\
&\leq K\sqrt{I_t(\boldsymbol{\gamma};A_t,\boldsymbol{Y}_t)}.
\end{aligned}$$

The second inequality is attained by noticing that $\sum_{k\in[N]}\mathbb{P}(k\in A_t)\leq K$, as at most K arms are played in each round. Note that this inequality typically show the benefits of information sharing among arms. Intuitively, with no feature sharing, we need to learn N independent meta parameters separately, and we gain mutual information for each arm-specific meta parameter only when the corresponding arm is pulled. However, with feature sharing, we keep gaining information for $\boldsymbol{\gamma}$, which leads to a lower regret for learning meta parameter.

□

F Experiment details

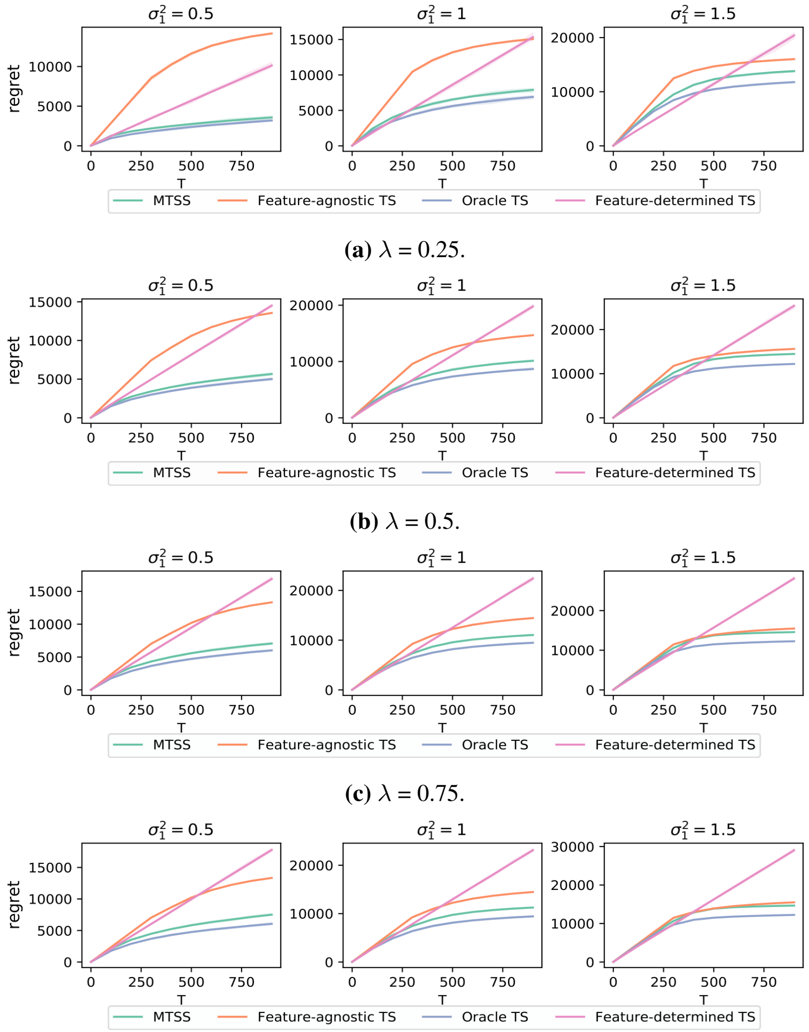
F.1 Robustness to model misspecification

To facilitate scalablity, we assume that $\theta_i|\mathbf{x}_i,\boldsymbol{\gamma}\sim g(\theta_i|\mathbf{x}_i,\boldsymbol{\gamma})$. When the model g is correctly specified, MTSS has shown superior theoretical and numerical performance.

Intuitively, as this model is used to construct a prior for the feature-agnostic model, as long as the learned priors provide reasonable information compared to the manually specified ones, this framework is still valuable. Indeed, related feature-agnostic TS algorithms typically enjoy prior-independent or instance-independent sublinear regrets [Wang and Chen, 2018, Perrault et al., 2020, Zhong et al., 2021].

In this section, we numerically study the impact of model misspecification on MTSS. When focusing on semi-bandits, the results under other problems are similar and therefore omitted. Specifically, instead of generating data according to $\theta_i\sim\mathcal{N}\big(\mathbf{x}_i^T\boldsymbol{\gamma},\sigma_1^2\big)$, we consider the data generation process $\theta_i\sim\mathcal{N}\big(\lambda cos(c_i\mathbf{x}_i^T\boldsymbol{\gamma})/c_i+(1-\lambda)\mathbf{x}_i^T\boldsymbol{\gamma},\sigma_1^2\big)$, where c is a normalization constant such that $c_i\mathbf{x}_i^T\boldsymbol{\gamma}\in[-\pi/2,\pi/2]$, and $\lambda\in[0,1]$ controls the degree of misspecification. When $\lambda=0$, we are considering the LMM; while when $\lambda=1$, the features provide few information through such a linear form.

In results reported in Figure 3, we observe that MTSS is fairly robust to model misspecifications. Although when λ and σ_1 increase, the advantage over feature-agnostic decreases, MTSS still always outperforms. Notably, MTSS always yield a nice sublinear regret unlike feature-determined TS, which further demonstrates the claimed robustness. We can even see that, perhaps surprisingly, when $\lambda=1$, MTSS still outperforms feature-agnostic TS. This is mainly due to that, with an intercept term in $\mathbf{x}_i$, our algorithm can at least learn $\mathbb{P}(\theta_i|\boldsymbol{\gamma})$ and enjoy the corresponding benefits. This is similar to the observations in Kveton et al. [2021] and Basu et al. [2021].



(a) $\lambda = 0.25$.

(b) $\lambda = 0.5$.

(c) $\lambda = 0.75$.

(d) $\lambda = 1.0$.

Figure 3: Robustness results. Shaded areas indicate the standard errors of the averages.

F.2 Model estimation with sparse features

Methodologically, the proposed framework is general and does not specify whether γ is sparse or not. The proposed framework can address the sparsity issue with minor modifications as a special case of our problem setting. Specifically, we can use the spike-and-slab prior for γ, which is popular in Bayesian sparse regression.

As an example, using combinatorial semi-bandits, we numerically demonstrated in Fig 4 that: when the number of non-zero parameters is fixed (i.e., d = 4), the regret does not scale quickly with the total number of parameters (i.e., P) involved in the model specification, with the approach proposed above. In contrast, the regret increases much quicker with the non-sparse regression (with Gaussian prior).

F.3 Experiment results with cold-start problems

In real-world applications, the set of items is typically not fixed. New items will be frequently introduced, and old items will be removed. Since there is no logged data for those newly-added items, such a challenge is typically referred to as the *cold-start* problem.

In this section, we compare the performance of various methods with the existence of the *cold-start* problem. We use semi-bandits as an example. Specifically, we set $L = 1000, T = 1000, K = 5, p = 5, \sigma_1 = \sigma_2 = 1$. We start with

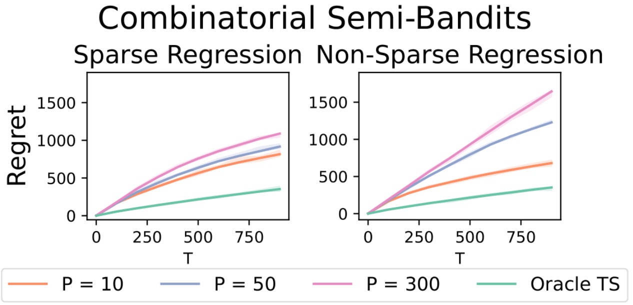


Figure 4: Experiment results for Sparse/Non-Sparse MTSS under Combinatorial Semi-Bandits.

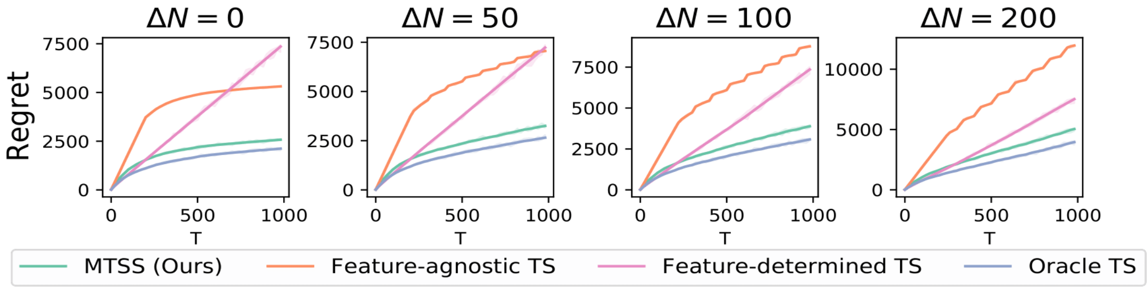


Figure 5: Experiment results for semi-bandits with the cold start problem.

$N = 1000$ items. The main difference with the experiments in the main text is that, every 100 time points, we will remove ΔN existing items and introduce ΔN new items. We vary the value of ΔN from 0 (no cold-start problem) to 200.

The experiment results can be found in Figure 5. As expected, in such a changing environment, all algorithms suffer a linear regret. The performance of feature-agnostic TS deteriorates significantly, as no information can be carried over to the new items. The difference between the regret of oracle-TS and MTSS is fairly stable, which implies that MTSS has learned the generalization function well and performs almost the same as oracle-TS eventually. MTSS consistently outperforms feature-agnostic TS and feature-determined TS.

F.4 Additional experiment results under other hyper-parameter settings

In this section, we present more simulation results under other combinations of L, K, d. See Figure 6 for details. Overall, the performance and conclusions are fairly consistent with the ones presented in the main text.

F.5 Additional experiment details

In this section, we first introduce how we evaluate the performance of the learning algorithms and the low-rank matrix factorization, which is widely used to construct features. Then, details for each real experiment are discussed.

Evaluation of Learning Algorithm. While the synthetic experiments compare the learning algorithms by Bayes Regret defined in the main context, here for the real experiment, we focus on the expected cumulative regret conditioned on the true $\boldsymbol{\theta}$, which is derived carefully from the dataset. Mathematically,

$$R(T, \boldsymbol{\theta}) = \sum_{t=1}^{T} \big[\max_{a \in \mathcal{A}} r(a, \boldsymbol{\theta}) - r(A_t, \boldsymbol{\theta}) \big].$$

Low-rank Matrix Factorization. Motivated by the collaborative filtering approach in recommender systems, low-rank factorization is widely used to construct the vectors of features. Suppose $A \in \mathcal{R}^{U \times N}$ includes the U observations of N items. Let $A \approx U \Sigma V^T$ be a rank-p truncated SVD of A, where $U \in \mathcal{R}^{U \times p}$, $\Sigma \in \mathcal{R}^{p \times p}$, and $V \in \mathcal{R}^{N \times p}$. Then the features of items are the rows of $V \Sigma$.

F.5.1 Cascading Bandits

Here, we use the data related to business and reviews from the Yelp Dataset Challenge, the usage license of which is described in Asghar [2016]. For our experiments, we extract $N = 3000$ restaurants with the most number of reviews and $U = 20K$ users writing the most number of reviews. Similar to Zong et al. [2016], we aim to maximize the probability of the user being attracted to at least one restaurant recommended. Following the experiment in Zong et al. [2016], we convert the review data to an observation matrix $W \in \{0,1\}^{U\times N}$, where each entry indicates if the user is attracted by the restaurant, by assuming that a restaurant will attract a user if the user reviewed the restaurants at least once before. After that, we split W into two distinct parts $W_{train} \in \{0,1\}^{\frac{U}{2}\times N}$ and $W_{test} \in \{0,1\}^{\frac{U}{2}\times N}$. While the W_{train} is used to extract the features of each restaurant, the W_{test} is used to evaluate the learning algorithms. Specifically, we applied the low-rank matrix factorization to W_{train} to derive the features of restaurants with $p = 10$. The final features are standardized in the experiment, and an intercept is considered, which leads to $d = 11$. Finally, the true parameter $\boldsymbol{\theta}$ is computed by taking the sample average of W_{test}, and the true parameter ϕ is derived appropriately from the W_{test} by analyzing its posterior distribution. For each round, the observation is randomly selected from W_{test}.

F.5.2 Semi-Bandits

Following the experiment setup in Wen et al. [2015], we use the Adult dataset, whose usage license is described in Dua and Graff [2017]. The Adult dataset includes features of $33K$ people. In our experiment, we focus on only $N = 3000$ people randomly selected. Our objective is to identify a set of $K = 20$ users among the 3000 people, including ten females and ten males, that are most likely to accept an advertisement. We considered $d = 4$ features including age, gender, whether the person works more than 40 hrs per week, and the length of education in years. Finally, we compute the true parameters from the dataset appropriately. First, we assume that the true expected acceptance probability (i.e., $\boldsymbol{\theta}$) depends on the user's income class. Specifically,

$$\theta_i = \begin{cases} .15, & \text{income} > \text{50K}. \\ .05, & \text{otherwise}. \end{cases}$$

Then, the true parameter σ_1 is learned by investigating the corresponding posterior distribution.

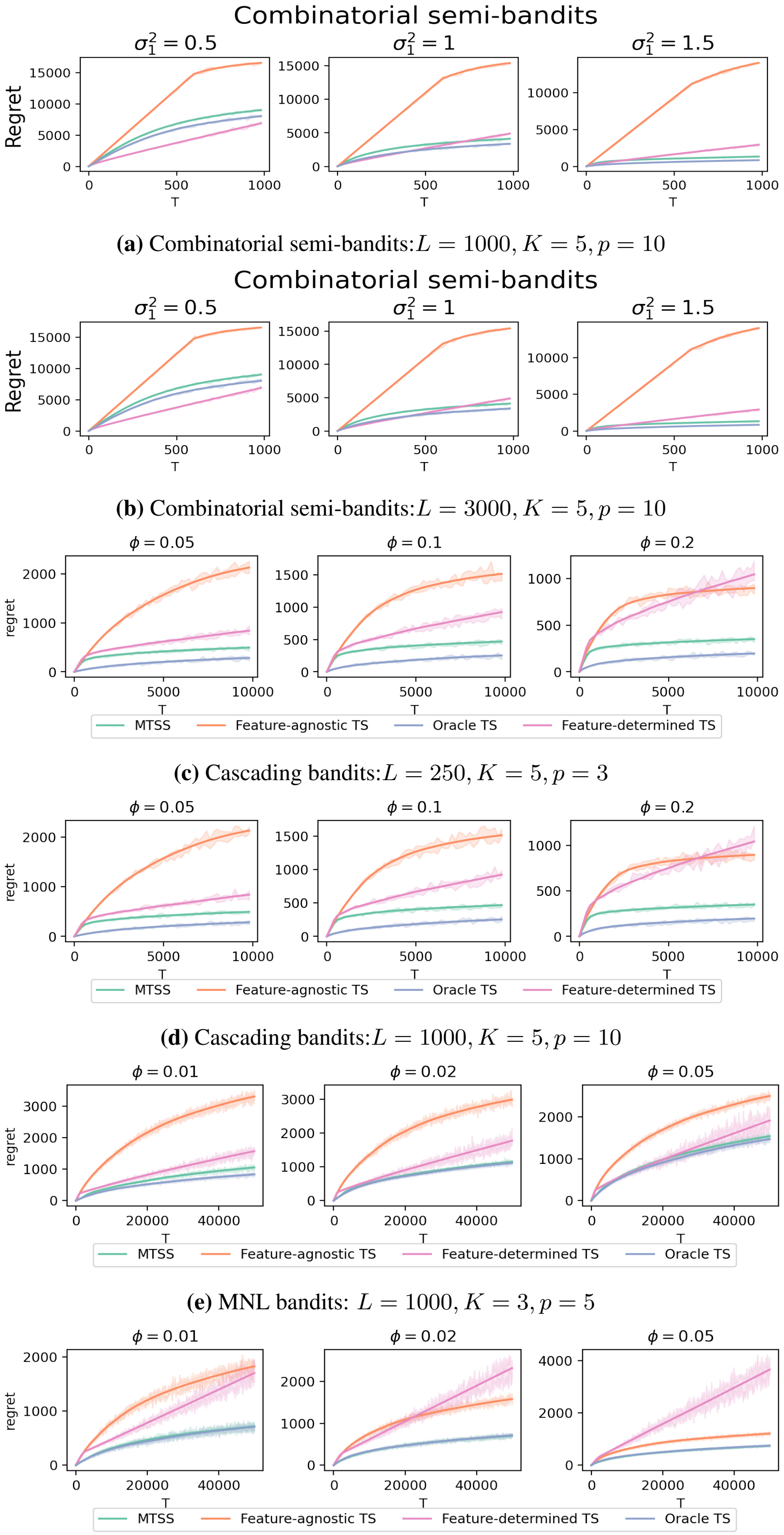
F.5.3 MNL Bandits

Following the experimental setup in Oh and Iyengar [2019], we use the dataset "MovieLens 1M" for our experiment, the usage license of which is described in Harper and Konstan [2015]. The dataset includes 1 million ratings of $6K$ movies from $U = 4K$ users. In our experiment, we use $N = 1000$ movies with the most ratings. While the range of ratings is from 1 to 5, we divide the ratings by 5 and consider it the utility of a movie to a user. Let the rating matrix be $W \in \mathcal{R}^{U\times N}$. We split W into equal-size training dataset $W_{train} \in \mathcal{R}^{\frac{U}{2}\times N}$, and test dataset $W_{test} \in \mathcal{R}^{\frac{U}{2}\times N}$. Since most ratings are not complete, as most users do not review all the selected movies, we first implement the low-rank matrix completion Keshavan et al. [2009] to fill the missing ratings in W_{train}. Similar to Oh and Iyengar [2019], we then apply the low-rank matrix factorization to the imputed W_{train} to construct the feature vector of each movie with $p = 5$. Then, we use the $L2$ normalization technique to normalize the features. Also, we consider including an intercept in the model. Therefore, $d = 6$ in the experiment. After that, we get the true mean utility v_i of each movie as the sample average of W_{test}, and the true parameter $\boldsymbol{\theta}$ is obtained directly. Finally, we learn the true parameter ϕ from W_{test} in the same way as before.

F.5.4 Real experiments with larger datasets

For a fair comparison, we chose the sample size N within the ranges used in related feature-determined papers [Oh and Iyengar, 2019, Wen et al., 2015, Zong et al., 2016], using the same datasets they used. However, we believe a more extensive dataset will help support the performance of the proposed methods even more.

Considering the limitation of large open datasets, in the following, we try our best to repeat the experiments with a larger size for all three problem instances, using the same set of datasets and similar pre-processing steps. Specifically, for the Yelp dataset, we increase N from 3000 to 8000 candidate restaurants; for the Adult dataset, we increase N from 3000 to 8000 individuals; for the MovieLens dataset, we increase N from 1000 to 2818 movies. The results showing in Fig 7 are similar to what we presented in the main context, with the proposed method consistently outperforming the baseline approaches.



(a) Combinatorial semi-bandits:$L = 1000, K = 5, p = 10$

(b) Combinatorial semi-bandits:$L = 3000, K = 5, p = 10$

(c) Cascading bandits:$L = 250, K = 5, p = 3$

(d) Cascading bandits:$L = 1000, K = 5, p = 10$

(e) MNL bandits: $L = 1000, K = 3, p = 5$

(f) MNL bandits: $L = 1000, K = 10, p = 3$

Figure 6: Simulation results under additional settings.

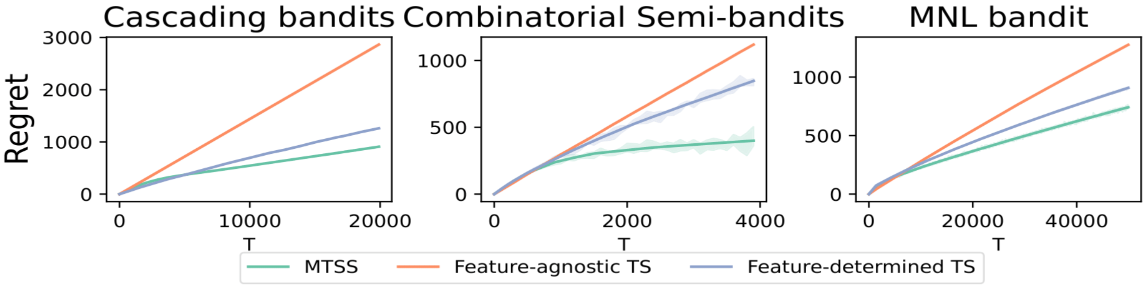


Figure 7: Experiment results for larger datasets.