Improved Bayes Regret Bounds for Multi-Task Hierarchical Bayesian Bandit Algorithms

Jiechao Guan, Hui Xiong

AI Thrust, The Hong Kong University of Science and Technology (Guangzhou), China

{jiechaoguan, xionghui}@hkust-gz.edu.cn

October 16, 2024

Presentation Outline

1. Background
    - Single-Task Bayesian Bandit
    - Multi-Task Bayesian Bandit
2. Algorithms
    - Hierarchical Thompson Sampling & Hierarchical BayesUCB
3. Improved Bayes Regret Bounds for Multi-Task Bandit
    - Near-Optimal Bayes Regret Bound for HierTS
    - Logarithmic Bayes Regret Bound for HierBayesUCB
4. Improved Bayes Regret Bounds for Multi-Task Semi-Bandit
    - Improved Bounds for HierTS and HierBayesUCB
5. Experiments
6. Conclusions

Single-Task Bandit

- A stochastic bandit problem is characterized by an unknown parameter θ with an action set $\mathcal{A}$. Each action $a \in \mathcal{A}$ under the bandit instance θ is associated with a reward distribution $\mathbb{P}(\cdot|a, \theta)$.
- The reward mean of action a under θ is denoted as $r(a;\theta) = \mathbb{E}_{Y\sim\mathbb{P}(\cdot|a;\theta)}[Y]$, and the optimal action under θ is denoted as $A_* = \arg\max_{a\in\mathcal{A}} r(a;\theta)$. In the stochastic linear bandit setting, the mean reward of action $a \in \mathcal{A}$ is $r(a, \theta) = a^\top\theta$.
- In Bayesian bandit problem, we further assume that the task parameter θ is independently and identically distributed (i.i.d.) according to a task parameter distribution $\mathbb{P}(\cdot|\mu_*)$, which is characterized by an unknown hyper-parameter μ_*.

Single-Task Semi-Bandit

- In this setting, the action set $\mathcal{A} = [K]$ is a set of finite items. $\mathscr{A} = \{A \subseteq \mathcal{A} : |A| \leq L\}$ is a family of subsets of $\mathcal{A}$ with up to L items, where $L \leq K$.
- $\mathbf{w} \in \mathbb{R}^K$ is a weight vector. The weight of a set $A \in \mathscr{A}$ is defined as $\sum_{a \in A} \mathbf{w}(a)$. We assume that the weights $\mathbf{w}$ are drawn i.i.d. from a distribution, and the mean weight is denoted as $\bar{\mathbf{w}} = \mathbb{E}[\mathbf{w}]$.
- We focus on the coherent case [1] which assumes that the agent knows a feature matrix $\Phi \in \mathbb{R}^{K \times d}$, such that $\bar{\mathbf{w}} = \Phi\theta$, where θ is the task parameter drawn from $\mathbb{P}(\cdot|\mu_*)$.
- The reward of a subset $A \in \mathscr{A}$ under the bandit instance θ is defined as $r(A; \theta) = \sum_{a \in A} (\Phi\theta)(a) = \sum_{a \in A} \langle \Phi_a, \theta \rangle$, where Φ_a is the transpose of the a-th row of matrix Φ. We further assume $\|\Phi_a\| \leq B, \forall a \in \mathcal{A}$.

Hierarchical Multi-Task Bayesian (Semi-)Bandit

- In this setting, the agent interacts with m tasks sequentially or concurrently. First, sample the hyper-parameter μ_* from a hyper-prior Q. Then, for each task $s \in [m]$, sample the task parameter $\theta_{s,*}$ independently from distribution $\mathbb{P}(\cdot|\mu_*)$.
- At round $t \geq 1$, the agent interacts with a set of tasks $\mathcal{S}_t \subseteq [m]$, takes a series of actions $A_t = (A_{s,t})_{s \in \mathcal{S}_t}$, and receives a series of rewards $Y_t = (Y_{s,t})_{s \in \mathcal{S}_t}$. In the bandit setting, $Y_{s,t} \sim \mathbb{P}(\cdot|A_{s,t}; \theta_{s,*})$ is a stochastic reward obtained by taking action $A_{s,t}$ in task $s \in \mathcal{S}_t$; in the semi-bandit setting, $Y_{s,t} = \{\hat{\mathbf{w}}_{s,t}(a)\}_{a \in A_{s,t}}$ is a series of stochastic rewards, where $\hat{\mathbf{w}}_{s,t} = \bar{\mathbf{w}}_s + \eta_{s,t}$, $\bar{\mathbf{w}}_s = \Phi\theta_{s,*}$, and $\eta_{s,t}$ is a K-dimensional random noise.
- The full hierarchical Bayesian bandit/semi-bandit model in the m-task learning setting is exhibited as follow for any $t \geq 1, s \in \mathcal{S}_t$:

$$\textbf{(1)}\ \mu_* \sim Q;\ \textbf{(2)}\ \theta_{s,*}|\mu_* \sim \mathbb{P}(\cdot|\mu_*), \forall s \in [m];\ \textbf{(3)}\ Y_{s,t}|A_{s,t}, \theta_{s,*} \sim \mathbb{P}(\cdot|A_{s,t}; \theta_{s,*}).$$

Multi-Task Bayes Regret

The goal of hierarchical Bayesian multi-task bandit/semi-bandit learning is to interact with m tasks efficiently and minimize the following cumulative *multi-task Bayes regret*:

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

where $A_{s,*} = \arg\max_{a\in\mathcal{A}} r(a;\theta_{s,*})$ is the optimal action for task $s \in [m]$ in the bandit setting, and $A_{s,*} \in \arg\max_{A\in\mathscr{A}} r(A;\theta_{s,*})$ is the optimal subset for task $s \in [m]$ in the semi-bandit setting.

Algorithm 1 Hierarchical Bayesian Algorithms for Multi-Task Linear Bandit Setting

```
Input: Hyper-prior Q
Initialize Q_1 ← Q
for t = 1, 2, ... do
    Sample hyper-parameter μ_t ~ Q_t
    Observe tasks S_t ⊆ [m]
    for s ∈ S_t do
        Option I (HierTS):
        Compute P_{s,t}(θ | μ_t) ∝ L_{s,t}(θ) P(θ | μ_t)
        Sample task parameter θ_{s,t} ~ P_{s,t}(· | μ_t)
        Take action A_{s,t} ← arg max_{a∈A} a^T θ_{s,t}
        Option II (HierBayesUCB):
        Set U_{t,s,a} = a^T μ̂_{s,t} + sqrt(2 log(1/δ)) ||a||_{Σ̂_{s,t}},
        for any a ∈ A
        Take action A_{s,t} ← arg max_{a∈A} U_{t,s,a}
        Observe reward Y_{s,t}
    end for
    Update Q_{t+1}
end for
```

Algorithm 2 Hierarchical Bayesian Algorithms for Multi-Task Combinatorial Semi-Bandit Setting

```
Input: Hyper-prior Q, features Φ ∈ R^{K×d}
Initialize Q_1 ← Q
for t = 1, 2, ... do
    Sample hyper-parameter μ_t ~ Q_t
    Observe tasks S_t ⊆ [m]
    for s ∈ S_t do
        Option I (HierTS):
        Compute P_{s,t}(θ | μ_t) ∝ L_{s,t}(θ) P(θ | μ_t)
        Sample task parameter θ_{s,t} ~ P_{s,t}(· | μ_t)
        Compute A_{s,t} = ORACLE(A, 𝒜, Φθ_{s,t})
        Option II (HierBayesUCB):
        Compute U_{t,s}(A) = Σ_{a∈A} (a^T μ̂_{s,t} + sqrt(2 log(1/δ)) ||a||_{Σ̂_{s,t}}), for all A ∈ 𝒜
        Compute A_{s,t} = arg max_{A∈𝒜} U_{t,s}(A)
        Chooose A_{s,t} and observe {ŵ_{s,t}(a)}_{a∈A_{s,t}}
    end for
    Update Q_{t+1}
end for
```

- Sampling $\theta_{s,t} \sim \mathbb{P}_{s,t}(\cdot|\mu_t)$ is equivalent to $\theta_{s,t} \sim \mathbb{P}(\theta_{s,*} = \theta|H_t)$
- $\hat{\mu}_{s,t}$ and $\hat{\Sigma}_{s,t}$ are the expectation and covariance of $\theta_{s,*} = \theta|H_t$.

Table 1: Different Bayes regret bounds for multi-task d-dimensional linear (or K-armed) bandit problem in the sequential setting. m is the number of tasks, n the number of iterations per task, $\mathcal{A}$ is the action set. **Bayes Regret Bound =Bound I + Bound II + Negligible Terms**, where **Bound I** is the regret bound for solving m tasks, **Bound II** the regret bound for learning hyper-parameter μ_*.

Bayes Regret Bound	$\lvert\mathcal{A}\rvert$	Bound I	Bound II
[25, ICML2021,Thm 3]	Finite	$O\left(m\sqrt{Kn\log n}\right)$	$O\left(n^2K\sqrt{m\log(n)\log(K)}\right)$
[7, NeurIPS2021, Thm 5]	Finite	$O\left(m\sqrt{dn(\log n)\log(n^2\lvert\mathcal{A}\rvert)}\right)$	$O\left(\sqrt{dmn(\log m)\log(n\lvert\mathcal{A}\rvert)}\right)$
[17, AISTAT2022, Thm 3]	Infinite	$O\left(md\sqrt{n\log(\frac{n}{d})\log(mn)}\right)$	$O\left(d\sqrt{mn\log(m)\log(mn)}\right)$
Our Theorem 5.1	Infinite	$O\left(md\sqrt{n\log(\frac{n}{d})}\right)$	$O\left(d\sqrt{mn\log(\frac{m}{d})}\right)$
Our Theorem 5.2	Finite	$O\left(md\log(\frac{n}{d})\log(mn)\right)$	$O\left(d\log(\frac{m}{d})\log(mn)\right)$

Near-Optimal Bayes Regret Bound for HierTS

Theorem 5.1 *(Near-Optimal Sequential Regret) Let $|\mathcal{S}_t| = 1$ for any round t. Then in the multi-task Gaussian linear bandit setting, the Bayes regret upper bound of HierTS is as follow:*

$$\mathcal{BR}(m,n) \leq d\sqrt{2mn}\sqrt{mc_1 \log\left(1+\frac{n}{d}\right) + c_2 \log\left(1+\frac{m\,\mathrm{Tr}(\Sigma_q\Sigma_0^{-1})}{d}\right)}.$$

- The term $md\sqrt{nc_1 \log(1+n/d)}$ represents the regret bound for solving m bandit tasks, whose parameters $\theta_{s,*}$ are drawn i.i.d. from the prior distribution $\mathcal{N}(\mu_*, \Sigma_0)$. Under this assumption, no task provides information for any other task, and hence this bound is linear in m. Similar observation was also pointed out by [2, 3, 4].
- The term $d\sqrt{mnc_2 \log\left(1+mtr(\Sigma_q\Sigma_0^{-1})/d\right)}$ represents the regret bound for learning the hyper-parameter μ_*.

Logarithmic Regret Bound for HierBayesUCB

Theorem 5.2 *(Logarithmic Sequential Regret of HierBayesUCB) Let $|\mathcal{S}_t| = 1$ for any round t, and the action set $\mathcal{A}$ is finite with $|\mathcal{A}| < \infty$. Then in the multi-task Gaussian linear bandit setting, for any $\delta \in (0,1)$, $\epsilon > 0$, the Bayes regret $\mathcal{BR}(m,n)$ of HierBayesUCB is upper bounded by*

$$mn\Big[\epsilon + 4B\delta\lambda_1^{\frac{1}{2}}(\Sigma_0+\Sigma_q)\big(d^{\frac{1}{2}} + \|\mu_q\|_{\hat{\Sigma}_{s,1}^{-1}}\big)|\mathcal{A}|\Big] + \mathbb{E}\Big[\frac{16d\log\frac{1}{\delta}}{\Delta_{\min}^{\epsilon}}\Big]\Big[mc_1\log\big(1+\frac{n}{d}\big) + c_2\log\big(1+\frac{m\operatorname{Tr}(\Sigma_q\Sigma_0^{-1})}{d}\big)\Big].$$

- If let $\delta = 1/(mn)$, $\epsilon = 1/(mn)$ and $\Delta_{\min} >> \epsilon$, the above sequential regret bound is of $O\big(\log(mn)(md\log(\frac{n}{d}) + d\log(\frac{m}{d}))\big)$.
- We can obtain sharper bounds by setting δ, ϵ as different values. For example, by setting $\delta = 1/n$, our regret bound becomes $O([mn\epsilon + m] + \frac{\log n}{\Delta_{\min}^{\epsilon}} m\log n)$, which is of order $O(m\log^2 n)$ if we set $\epsilon = 1/(mn)$ and the gap $\Delta_{\min} >> \epsilon$ is large.

Table 2: Different Bayes regret bounds for multi-task semi-bandit problem. **Bayes Regret Bound =Bound I + Bound II + Negligible Terms**. m is the number of tasks, n the number of iterations per task, K the size of action set, L the number of pulled actions at each round ($1 \le L \le K$). **Bound I** is the regret bound for solving m tasks, **Bound II** the regret bound for learning hyper-parameter μ_*.

Bayes Regret Bound	$\mathcal{A}$	**Bound I**	**Bound II**
[7, Theorem 6]	$[K]$	$O\left(m\sqrt{nKL\log n \log(nK)}\right)$	$O\left(\sqrt{mnKL\log m \log(nK)}\right)$
Our Theorem 5.4	$[K]$	$O\left(m\sqrt{nL\log(nL)\log(nK)}\right)$	$O\left(L^{\frac{3}{2}}\sqrt{mn\log m \log(nK)}\right)$
Our Theorem 5.5	$[K]$	$O\left(mL\log(nL)\log(mnK)\right)$	$O\left(L^3\log(m)\log(mnK)\right)$

Bayes Regret Bounds for Semi-Bandit

Theorem 5.4 *Let $|\mathcal{S}_t| = 1$ for any $t \geq 1$. Let $c \geq \sqrt{2\ln\left(\frac{nKB\lambda_1(\Sigma_0)}{\sqrt{2\pi}}\right)}$, then in the multi-task Gaussian semi-bandit setting, the Bayes regret upper bound of combinatorial HierTS is:*

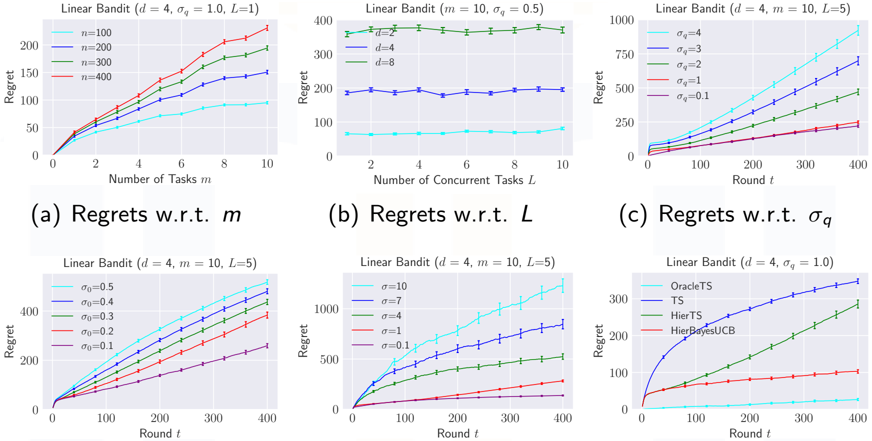
$$\mathcal{BR}(m,n) \leq m + c\sqrt{mnL}\sqrt{2c_1 m \log\left(1+\frac{nL}{d}\right) + 2c_4 Ld\log\left(1+\frac{m\,\mathrm{Tr}(\Sigma_0^{-1}\Sigma_q)}{d}\right)}.$$

- HierTS obtains $O(m\sqrt{n}\log n)$ Bayes regret for semi-bandit.

Theorem 5.5 *Let $|\mathcal{S}_t| = 1$ for any $t \geq 1$. Then for any $\epsilon > 0, \delta \in (0,1)$, in the multi-task Gaussian semi-bandit setting, the Bayes regret $\mathcal{BR}(m,n)$ of combinatorial HierBayesUCB is bounded by*

$$mn\left[\epsilon + 4LBK\delta\lambda_1^{\frac{1}{2}}(\Sigma_0+\Sigma_q)(d^{\frac{1}{2}} + \|\mu_q\|_{\hat{\Sigma}_{s,1}^{-1}})\right] + \mathbb{E}\left[\frac{8L\log\frac{1}{\delta}}{\Delta_{\min}^{\epsilon}}\right]\left[2c_1 m\log\left(1+\frac{nL}{d}\right) + 2c_4 Ld\log\left(1+\frac{m\,\mathrm{Tr}(\Sigma_0^{-1}\Sigma_q)}{d}\right)\right]$$

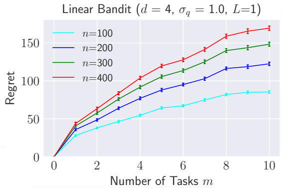
- HierBayesUCB obtains $O(m\log(mn)\log n)$ Bayes regret in for semi-bandit.



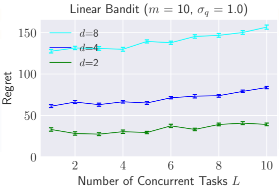
(a) Regrets w.r.t. m (b) Regrets w.r.t. L (c) Regrets w.r.t. σ_q

(d) Regrets w.r.t. σ_0 (e) Regrets w.r.t. σ (f) Regrets of algorithms

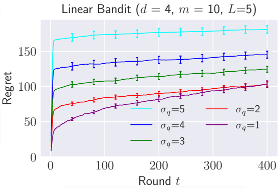
Figure 1: Regrets of HierTS w.r.t. different hyper-parameters.



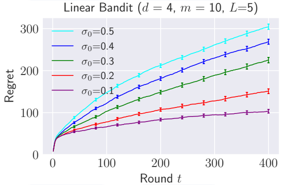
(a) Regrets w.r.t. m



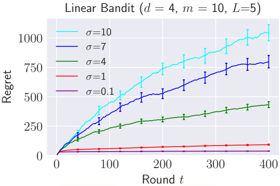
(b) Regrets w.r.t. L



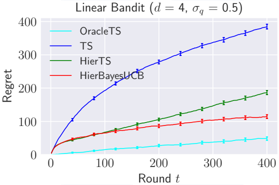
(c) Regrets w.r.t. σ_q



(d) Regrets w.r.t. σ_0



(e) Regrets w.r.t. σ



(f) Regrets of algorithms

Figure 2: Regrets of HierBayesUCB w.r.t. different hyper-parameters.

Conclusions

Our theoretical contributions are four-fold:

- In the case of infinite action set, we provide a tighter Bayes regret bound $O(m\sqrt{n\log n})$ for HierTS. This bound improves the latest result by a factor of $O(\sqrt{\log(mn)})$.
- In the case of finite action set, we propose a novel HierBayesUCB algorithm, and provide gap-dependent logarithmic Bayes regret bound $O(m\log(mn)\log n)$ for it.
- We generalize the above regret bounds for linear bandit from sequential setting to the more challenging concurrent setting.
- We extend both HierTS and HierBayesUCB algorithms to the more general multi-task combinatorial semi-bandit setting and derive improved Bayes regret bounds.

Thanks!

References

[1] Zheng Wen and Benjamin Van Roy. "Efficient Exploration and Value Function Generalization in Deterministic Systems". In: *NeurIPS*. 2013, pp. 3021–3029.

[2] Branislav Kveton et al. "Meta-Thompson Sampling". In: *ICML*. 2021, pp. 5884–5893.

[3] Soumya Basu et al. "No Regrets for Learning the Prior in Bandits". In: *NeurIPS*. 2021, pp. 28029–28041.

[4] Joey Hong et al. "Hierarchical Bayesian Bandits". In: *AISTATS*. 2022, pp. 7724–7741.