Meta-Thompson Sampling

Branislav Kveton [1] **Mikhail Konobeev** [2] **Manzil Zaheer** [1] **Chih-wei Hsu** [1] **Martin Mladenov** [1] **Craig Boutilier** [1] **Csaba Szepesvári** [3 2]

Abstract

Efficient exploration in bandits is a fundamental online learning problem. We propose a variant of Thompson sampling that learns to explore better as it interacts with bandit instances drawn from an unknown prior. The algorithm meta-learns the prior and thus we call it `MetaTS`. We propose several efficient implementations of `MetaTS` and analyze it in Gaussian bandits. Our analysis shows the benefit of meta-learning and is of a broader interest, because we derive a novel prior-dependent Bayes regret bound for Thompson sampling. Our theory is complemented by empirical evaluation, which shows that `MetaTS` quickly adapts to the unknown prior.

1. Introduction

A *stochastic bandit* (Lai & Robbins, 1985; Auer et al., 2002; Lattimore & Szepesvari, 2019) is an online learning problem where a *learning agent* sequentially interacts with an environment over n rounds. In each round, the agent pulls an *arm* and then receives the arm's *stochastic reward*. The agent aims to maximize its expected cumulative reward over n rounds. It does not know the mean rewards of the arms *a priori*, so must learn them by pulling the arms. This induces the well-known *exploration-exploitation trade-off*: *explore*, and learn more about an arm; or *exploit*, and pull the arm with the highest estimated reward. In a clinical trial, the arm might be a treatment and its reward is the outcome of that treatment for a patient.

Bandit algorithms are typically designed to have low regret for some problem class of interest to the algorithm designer (Lattimore & Szepesvari, 2019). In practice, however, the problem class may not be perfectly specified at the time of the design. For instance, consider applying *Thompson sampling (TS)* (Thompson, 1933; Chapelle & Li, 2012; Agrawal & Goyal, 2012; Russo et al., 2018) to a 2-armed bandit in which the prior distribution over mean arm rewards, a vital part of TS, is unknown. While the prior is unknown, the designer may know that it is one of two possible priors where either arm 1 or arm 2 is optimal with high probability. If the agent could learn which of the two priors has been realized, for instance by interacting repeatedly with bandit instances drawn from that prior, it could adapt its exploration strategy to the realized prior, and thereby incur much lower regret than would be possible without this adaptation.

We formalize this learning problem as follows. A learning agent sequentially interacts with m bandit instances. Each interaction has n rounds and we refer to it as a *task*. The instances share a common structure, namely that their mean arm rewards are drawn from an unknown *instance prior* P_*. While P_* is not known, we assume that it is sampled from a *meta-prior* Q, which the agent knows with certainty. The goal of the agent is to minimize the regret in each sampled instance almost as well as if it knew P_*. This is achieved by adapting to P_* through interactions with the instances. This is a form of *meta-learning* (Thrun, 1996; 1998; Baxter, 1998; 2000), where the agent learns to act from interactions with bandit instances.

We make the following contributions. First, we formalize the problem of Bayes regret minimization where the prior P_* is unknown, and is learned by interactions with bandit instances sampled from it in m tasks. Second, we propose `MetaTS`, a *meta-Thompson sampling* algorithm that solves this problem. `MetaTS` maintains a distribution over the unknown P_* in each task, which we call a *meta-posterior* Q_s, and acts optimistically with respect to it. More specifically, in task s, it samples an estimate of P_* as $P_s \sim Q_s$ and then runs TS with prior P_s for n rounds. We show how to implement `MetaTS` efficiently in Bernoulli and Gaussian bandits. In addition, we bound its Bayes regret in Gaussian bandits. Our analysis is conservative because it relies only on a single pull of each arm in each task. Nevertheless, it yields an improved regret bound due to adapting to P_*. The analysis is of broader interest, as we derive a novel *prior-dependent* upper bound on the Bayes regret of TS. Our theoretical results are complemented by synthetic experiments, which show that `MetaTS` adapts quickly to the unknown prior P_*, and its regret is comparable to that of TS with a known P_*.

[1] Google Research [2] University of Alberta [3] DeepMind. Correspondence to: Branislav Kveton <bkveton@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

2. Setting

We start with introducing our notation. The set $\{1, \dots, n\}$ is denoted by $[n]$. The indicator $\mathbb{1}\{E\}$ denotes that event E occurs. The i-th entry of vector v is denoted by v_i. Sometimes we write $v(i)$ to avoid clutter. A diagonal matrix with entries v is denoted by $\mathrm{diag}\,(v)$. We write $\tilde{O}$ for the big-O notation up to polylogarithmic factors.

Our setting is defined as follows. We have K arms, where a bandit *problem instance* is a vector of arm means $\theta \in \mathbb{R}^K$. The agent sequentially interacts with m bandit instances, which we index by $s \in [m]$. We refer to each interaction as a *task*. At the beginning of task $s \in [m]$, an instance $\theta_{s,*}$ is sampled i.i.d. from an *instance prior distribution* P_*. The agent interacts with $\theta_{s,*}$ for n rounds. In round $t \in [n]$, it pulls one arm and observes a stochastic realization of its reward. We denote the pulled arm in round t of task s by $A_{s,t} \in [K]$, the stochastic rewards of all arms in round t of task s by $Y_{s,t} \in \mathbb{R}^K$, and the reward of arm $i \in [K]$ by $Y_{s,t}(i)$. The result of the interactions in task s is *history*

$$H_s = (A_{s,1}, Y_{s,1}(A_{s,1}), \dots, A_{s,n}, Y_{s,n}(A_{s,n}))\,.$$

We denote by $H_{1:s} = H_1 \oplus \cdots \oplus H_s$ a concatenated vector of the histories in tasks 1 to s. We assume that the realized rewards $Y_{s,t}$ are i.i.d. with respect to both s and t, and that their means are $\mathbb{E}\left[Y_{s,t} \,|\, \theta_{s,*} = \theta\right] = \theta$. For now, we need not assume that the reward noise is sub-Gaussian; but we do adopt this in our analysis (Section 4).

The n-round *Bayes regret* of a learning agent or algorithm over m tasks with instance prior P_* is

$$R(m, n; P_*) = \sum_{s=1}^{m} \mathbb{E}\left[\sum_{t=1}^{n} \theta_{s,*}(A_{s,*}) - \theta_{s,*}(A_{s,t}) \,\middle|\, P_*\right]\,,$$

where $A_{s,*} = \arg\max_{i \in [K]} \theta_{s,*}(i)$ is the *optimal arm* in the problem instance $\theta_{s,*}$ in task $s \in [m]$. The above expectation is over problem instances $\theta_{s,*}$ sampled from P_*, their realized rewards, and pulled arms.

We note that $R(1, n; P_*)$ is the standard definition of the n-round Bayes regret in a K-armed bandit (Russo & Van Roy, 2014), and that it is $\tilde{O}(\sqrt{Kn})$ for Thomson sampling with prior P_*. Since all bandit instances $\theta_{s,*}$ are drawn i.i.d. from the same P_*, they provide no additional information about each other. Thus, the regret of TS with prior P_* in m such instances is $\tilde{O}(m\sqrt{Kn})$. We validate this dependence empirically in Section 5.

Note that Thompson sampling requires P_* as an input. In this work, we try to attain the same regret *without assuming that P_* is known*. We formalize this problem in a Bayesian fashion. In particular, we assume the availability of a prior distribution Q over problem instance priors, and that $P_* \sim Q$. We refer to Q as a *meta-prior* since it is a prior over priors. In Bayesian statistics, this would also be known as a hyper-prior (Gelman et al., 2013). The agent knows Q but not P_*. We try to learn P_* from sequential interactions with instances $\theta_{s,*}$, which are drawn i.i.d. from P_* in each task. Note that $\theta_{s,*}$ is also unknown. The agent only observes its noisy realizations $Y_{s,t}$. We visualize relations of Q, P_*, $\theta_{s,*}$, and $Y_{s,t}$ in a graphical model in Figure 1.

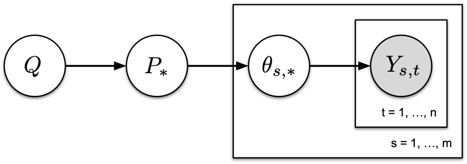

Figure 1. Graphical model of our bandit environment.

One motivating example for using a meta-prior for exploration arises in recommender systems, in which exploration is used to assess the latent interests of users for different items, such as movies. In this case, each user can be treated as a bandit instance where the items are arms. A standard prior over user latent interests could readily be used by TS (Hong et al., 2020). However, in many cases, the algorithm designer may be uncertain about its true form. For instance, the designer may believe that most users have strong but noisy affinity for items in exactly one of several classes, but it is unclear which. Our work can be viewed as formalizing the problem of learning such a prior over user interests, which could be used to start exploring the preferences of "cold-start" users.

3. Meta-Thompson Sampling

In this section, we present our approach to meta-learning in TS. We provide a general description in Section 3.1. In Sections 3.2 and 3.3, we implement it in Bernoulli bandits with a categorical meta-prior and Gaussian bandits with a Gaussian meta-prior, respectively. In Section 3.4, we justify our approach beyond these specific instances.

3.1. Algorithm MetaTS

Thompson sampling (Thompson, 1933; Chapelle & Li, 2012; Agrawal & Goyal, 2012; Russo et al., 2018) is arguably the most popular and practical bandit algorithm. TS is parameterized by a prior, which is specified by the algorithm designer. In this work, we study a more general setting where the designer can model uncertainty over an unknown prior P_* using a meta-prior Q. Our proposed algorithm meta-learns P_* from sequential interactions with bandit instances drawn i.i.d. from P_*. Therefore, we call it *meta-Thompson sampling* (MetaTS). In this subsection, we present MetaTS under the assumption that sample spaces are discrete. This eases exposition and guarantees that all

Algorithm 1 MetaTS: Meta-learning Thompson sampling.

```
Inputs: Meta-prior $Q$

$Q_1 \leftarrow Q$
for $s = 1, \dots, m$ do
    Sample $P_s \sim Q_s$
    Apply Thompson sampling with prior $P_s$ to
    problem instance $\theta_{s,*} \sim P_*$ for $n$ rounds
    Update meta-posterior $Q_{s+1}$, as defined in (1)
```

conditional expectations are properly defined. We treat this topic more rigorously in Section 3.4.

MetaTS is a variant of TS that models uncertainty over the instance prior distribution P_*. This uncertainty is captured by a meta-posterior Q_s, a distribution over possible instance priors. We denote the *meta-posterior* in task s by Q_s, and assume that each Q_s belongs to the same family as Q. By definition, $Q_1 = Q$ is the *meta-prior*. MetaTS samples the *instance prior distribution* P_s in task s from Q_s. Then it applies TS with the sampled prior P_s to the bandit instance $\theta_{s,*}$ in task s for n rounds. Once the task is complete, it updates the meta-posterior in a standard Bayesian fashion

$$
\begin{aligned}
& Q_{s+1}(P) \qquad (1)\\
& \propto \mathbb{P}\left(H_{1:s} \,|\, P_* = P\right) Q(P) \\
& = \mathbb{P}\left(H_s \,|\, P_* = P\right) \prod_{\ell=1}^{s-1} \mathbb{P}\left(H_\ell \,|\, P_* = P\right) Q(P) \\
& = \mathbb{P}\left(H_s \,|\, P_* = P\right) Q_s(P) \\
& = \int_\theta \mathbb{P}\left(H_s \,|\, \theta_{s,*} = \theta\right) \mathbb{P}\left(\theta_{s,*} = \theta \,|\, P_* = P\right) \mathrm{d}\theta \, Q_s(P) \,,
\end{aligned}
$$

where $\mathbb{P}\left(H_s \,|\, P_* = P\right)$ and $\mathbb{P}\left(H_s \,|\, \theta_{s,*} = \theta\right)$ are probabilities of observations in task s given that the instance prior is P and the problem instance is θ, respectively. A rigorous justification of this update is given in Section 3.4. Specific instances of this update are in Sections 3.2 and 3.3.

The pseudocode for MetaTS is presented in Algorithm 1. The algorithm is simple, natural, and general; but has two potential shortcomings. First, it is unclear if it can be implemented efficiently. To address this, we develop efficient implementations for both Bernoulli and Gaussian bandits in Sections 3.2 and 3.3, respectively. Second, it is unclear whether MetaTS explores enough. Ideally, it should *learn to* perform as well as TS with the true prior P_*. Intuitively, we expect this since the meta-posterior samples $P_s \sim Q_s$ should vary significantly in the direction of high variance in Q_s, which represents high uncertainty that can be reduced by exploring. We confirm this in Section 4, where MetaTS is analyzed in Gaussian bandits.

3.2. Bernoulli Bandit with a Categorical Meta-Prior

Bernoulli Thompson sampling was the first instance of TS that was analyzed (Agrawal & Goyal, 2012). In this section, we apply MetaTS to this problem class.

We consider a Bernoulli bandit with K arms that is parameterized by arm means $\theta \in [0, 1]^K$. The reward of arm i in instance θ is drawn i.i.d. from $\mathrm{Ber}(\theta_i)$. To model uncertainty in the prior, we assume access to L potential instance priors $\mathcal{P} = \left\{P^{(j)}\right\}_{j=1}^L$. Each prior $P^{(j)}$ is factored across the arms as

$$
\begin{aligned}
P^{(j)}(\theta) & = \prod_{i=1}^K \mathrm{Beta}(\theta_i; \alpha_{i,j}, \beta_{i,j}) \\
& = \prod_{i=1}^K \frac{\Gamma(\alpha_{i,j} + \beta_{i,j})}{\Gamma(\alpha_{i,j})\Gamma(\beta_{i,j})} \theta_i^{\alpha_{i,j}-1} (1 - \theta_i)^{\beta_{i,j}-1}
\end{aligned}
$$

for some fixed $(\alpha_{i,j})_{i=1}^K$ and $(\beta_{i,j})_{i=1}^K$. The *meta-prior* is a categorical distribution over L classes of tasks. That is,

$$
Q(j) = \mathrm{Cat}(j; w) = w_j
$$

for $w \in \Delta_{L-1}$, where w is a vector of initial beliefs into each instance prior and Δ_{L-1} is the L-dimensional simplex. The tasks are generated as follows. First, the instance prior is set as $P_* = P^{(j_*)}$ where $j_* \sim Q$. Then, in each task s, a Bernoulli bandit instance is sampled as $\theta_{s,*} \sim P_*$.

MetaTS is implemented as follows. The meta-posterior in task s is

$$
Q_s(j) = \mathrm{Cat}(j; \hat{w}_s) = \hat{w}_{s,j} \,,
$$

where $\hat{w}_s \in \Delta_{L-1}$ is a vector of posterior beliefs into each instance prior. The instance prior in task s is $P_s = P^{(j_s)}$ where $j_s \sim Q_s$. After interacting with bandit instance $\theta_{s,*}$, the meta-posterior is updated using $Q_{s+1}(j) \propto f(j) \, Q_s(j)$, where

$$
\begin{aligned}
f(j) & = \int_\theta \mathbb{P}\left(H_s \,|\, \theta_{s,*} = \theta\right) \mathbb{P}\left(\theta_{s,*} = \theta \,|\, j_* = j\right) \mathrm{d}\theta \\
& = \prod_{i=1}^K \frac{\Gamma(\alpha_{i,j} + \beta_{i,j})}{\Gamma(\alpha_{i,j})\Gamma(\beta_{i,j})} \times \\
& \qquad \int_{\theta_i} \theta_i^{\alpha_{i,j} + N_{i,s}^+ - 1} (1 - \theta_i)^{\beta_{i,j} + N_{i,s}^- - 1} \, \mathrm{d}\theta_i \\
& = \prod_{i=1}^K \frac{\Gamma(\alpha_{i,j} + \beta_{i,j}) \Gamma(\alpha_{i,j} + N_{i,s}^+) \Gamma(\beta_{i,j} + N_{i,s}^-)}{\Gamma(\alpha_{i,j}) \Gamma(\beta_{i,j}) \Gamma(\alpha_{i,j} + \beta_{i,j} + T_{i,s})} \,.
\end{aligned}
$$

Here $\mathcal{A}_{i,s} = \{t \in [n] : A_{s,t} = i\}$ is the set of rounds where arm i is pulled in task s and $T_{i,s} = |\mathcal{A}_{i,s}|$ is the number of these rounds. In addition, $N_{i,s}^+ = \sum_{t \in \mathcal{A}_{i,s}} Y_{s,t}(i)$ denotes the number of positive observations of arm i and $N_{i,s}^- = T_{i,s} - N_{i,s}^+$ is the number of its negative observations.

The above derivation can be generalized in a straightforward fashion to any categorical meta-prior whose instance priors $P^{(j)}$ lie in some exponential family.

3.3. Gaussian Bandit with a Gaussian Meta-Prior

Gaussian distributions have many properties that allow for tractable analysis, such as that the posterior variance is independent of observations, which we exploit in Section 4. In this section, we present a computationally-efficient implementation for this problem class.

We consider a Gaussian bandit with K arms that is parameterized by arm means $\theta \in \mathbb{R}^K$. The reward of arm i in instance θ is drawn i.i.d. from $\mathcal{N}(\theta_i, \sigma^2)$. We have a continuum of instance priors, parameterized by a vector of means $\mu \in \mathbb{R}^K$ and defined as $P(\theta) = \mathcal{N}(\theta; \mu, \sigma_0^2 I_K)$. The noise σ_0 is fixed. The *meta-prior* is a Gaussian distribution over instance prior means $Q(\mu) = \mathcal{N}(\mu; \mathbf{0}, \sigma_q^2 I_K)$, where σ_q is assumed to be known. The tasks are generated as follows. First, the instance prior is set as $P_* = \mathcal{N}(\mu_*, \sigma_0^2 I_K)$ where $\mu_* \sim Q$. Then, in each task s, a Gaussian bandit instance is sampled as $\theta_{s,*} \sim P_*$.

`MetaTS` is implemented as follows. The meta-posterior in task s is

$$Q_s(\mu) = \mathcal{N}(\mu; \hat{\mu}_s, \hat{\Sigma}_s)\,,$$

where $\hat{\mu}_s \in \mathbb{R}^K$ is an estimate of μ_* and $\hat{\Sigma}_s \in \mathbb{R}^{K \times K}$ is a diagonal covariance matrix. The instance prior in task s is $P_s(\theta) = \mathcal{N}(\theta; \tilde{\mu}_s, \sigma_0^2 I_K)$ where $\tilde{\mu}_s \sim Q_s$. After interacting with bandit instance $\theta_{s,*}$, the meta-posterior is updated as $Q_{s+1}(\mu) \propto f(\mu)\, Q_s(\mu)$, where

$$\begin{aligned} & f(\mu) \\ & = \int_\theta \mathbb{P}\left(H_s \,|\, \theta_{s,*} = \theta\right) \mathbb{P}\left(\theta_{s,*} = \theta \,|\, \mu_* = \mu\right) \mathrm{d}\theta \\ & = \prod_{i=1}^K \int_{\theta_i} \left[\prod_{t \in \mathcal{A}_{i,s}} \mathcal{N}(Y_{s,t}(i); \theta_i, \sigma^2)\right] \mathcal{N}(\theta_i; \mu_i, \sigma_0^2)\, \mathrm{d}\theta_i\,. \end{aligned}$$

Here $\mathcal{A}_{i,s} = \{t \in [n] : A_{s,t} = i\}$ is the set of rounds where arm i is pulled in task s and $T_{i,s} = |\mathcal{A}_{i,s}|$ is the number of such rounds, as in Section 3.2.

Since $\hat{\Sigma}_s = \operatorname{diag}\left(\hat{\sigma}_s^2\right)$ is a diagonal covariance matrix, it is fully characterized by a vector of individual arm variances $\hat{\sigma}_s^2 \in \mathbb{R}^K$. The parameters $\hat{\mu}_s$ and $\hat{\sigma}_s^2$ are updated, based on Lemma 7 in Appendix B, as

$$\begin{aligned} \hat{\mu}_{s+1,i} &= \hat{\sigma}_{s+1,i}^2 \left(\frac{\hat{\mu}_{s,i}}{\hat{\sigma}_{s,i}^2} + \frac{T_{i,s}}{T_{i,s}\sigma_0^2 + \sigma^2} \frac{\sum_{t \in \mathcal{A}_{i,s}} Y_{s,t}(i)}{T_{i,s}}\right)\,, \\ \hat{\sigma}_{s+1,i}^{-2} &= \hat{\sigma}_{s,i}^{-2} + \frac{T_{i,s}}{T_{i,s}\sigma_0^2 + \sigma^2}\,. \end{aligned}$$

This update can be also derived using (17) in Appendix D, when all covariance matrices are assumed to be diagonal.

The above update has a very nice interpretation. The posterior mean $\hat{\mu}_{s+1,i}$ of arm i is a weighted sum of the mean reward estimate of arm i in task s and the earlier posterior mean $\hat{\mu}_{s,i}$. The weight of the estimate depends on how good it is. Specifically, it varies from $1/(\sigma_{0,i}^2 + \sigma^2)$, when arm i is pulled only once, to $1/\sigma_{0,i}^2$, when $T_{i,s} \to \infty$. This is the minimum amount of uncertainty that cannot be reduced by more pulls, due to the fact that $\theta_{s,*}$ is a single observation of the unknown μ_* with covariance $\sigma_0^2 I_K$.

3.4. Measure-Theoretic View and the General Case

We now present a more general measure-theoretic specification of our meta-bandit setting. Let $\mathcal{Z}$ be the set of outcomes for the hidden variable Z that is sampled from a meta-prior and $\sigma(\mathcal{Z})$ be the σ-algebra over this set. Similarly, let Θ be the set of possible bandit environments $\theta \in \Theta$ and $\sigma(\Theta)$ be the σ-algebra over this set. While in this work we focus on environments characterized only by their mean reward vectors, this parameterization could be more general, and for example include the variance of mean reward vectors. The formal definition of a K-armed Bayesian meta-bandit is as follows.

Definition 1. *A K-armed Bayesian meta-bandit is a tuple $\mathcal{B} = (\mathcal{Z}, \sigma(\mathcal{Z}), Q, \Theta, \sigma(\Theta), P, \rho)$, where $(\mathcal{Z}, \sigma(\mathcal{Z}))$ is a measurable space; the* meta-prior *Q is a probability measure over $(\mathcal{Z}, \sigma(\mathcal{Z}))$; the* prior *$P$ is a probability kernel from $(\mathcal{Z}, \sigma(\mathcal{Z}))$ to $(\Theta, \sigma(\Theta))$; and $\rho = (\rho_{\theta,i} : \theta \in \Theta, i \in [K])$ is a probability kernel from $\Theta \times [K]$ to $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, where $\mathfrak{B}(\mathbb{R})$ is the Borel σ-algebra of $\mathbb{R}$ and $\rho_{\theta,i}$ is the reward distribution associated with arm i in bandit θ.*

We use lowercase letters to denote realizations of random variables. Let P_z be a distribution of bandit instances under $Z = z$. We assume that a new environment $\theta \in \Theta$ is sampled from the same measure P_z at the beginning of each task with the same realization of hidden variable Z sampled from the meta-prior Q beforehand.

A bandit algorithm consists of kernels $\pi_{s,t}$ that take as input a history of interactions consisting of the pulled arms and observed rewards up to round t in task s, and output a probability measure over the arms. A bandit algorithm is connected with a Bayesian meta-bandit environment $\mathcal{B}$ to produce a sequence of chosen arms and observed rewards. Formally, let $\Omega_{s,t} = ([K] \times \mathbb{R})^{(s-1)n+t-1} \subset \mathbb{R}^{2((s-1)n+t-1)}$ for each $t \in [n]$ and $s \in [m]$. Then a bandit algorithm or policy is a tuple $\pi = (\pi_{s,t})_{s,t=1}^{m,n}$ such that each $\pi_{s,t}$ is a kernel from $(\Omega_{s,t}, \mathfrak{B}(\mathbb{R}^{2((s-1)n+t-1)}))$ to $([K], 2^{[K]})$ that interacts with a Bayesian meta-bandit $\mathcal{B}$ over m tasks, each lasting n rounds and producing a sequence of random variables

$$\begin{aligned} & A_{1,1}, X_{1,1}, \dots, A_{1,n}, X_{1,n}\,, \dots, \\ & A_{m,1}, X_{m,1} \dots, A_{m,n}, X_{m,n}\,, \end{aligned}$$

where $X_{s,t} = Y_{s,t}(A_{s,t})$ is the reward in round t of task s. The probability measure over these variables, $\mathbb{P}_{z,\theta_1,\dots,\theta_m,\pi}$, is guaranteed to exist by the Ionescu-Tulcea theorem (Tulcea, 1949). Furthermore, the conditional probabilities of transitions of this measure are equal to the kernels

$$\begin{aligned} & \mathbb{P}\left(\theta_s \in \cdot \mid z, \theta_1, \dots, \theta_{s-1}, H_{1:s-1}\right) = P_z(\theta_s \in \cdot)\,, \\ & \mathbb{P}\left(A_{s,t} \in \cdot \mid z, \theta_1, \dots, \theta_s, H_{1:s-1}, H_{s,t}\right) = \\ & \qquad \pi_{s,t}(A_{s,t} \in \cdot \mid H_{1:s-1}, H_{s,t})\,, \\ & \mathbb{P}\left(X_{s,t} \in \cdot \mid z, \theta_1, \dots, \theta_s, H_{1:s-1}, H_{s,t}, A_{s,t}\right) = \\ & \qquad \rho_{\theta_s, A_{s,t}}(X_{s,t} \in \cdot)\,, \end{aligned}$$

where $H_{s,t} = (X_{s,\ell}, A_{s,\ell})_{\ell=1}^{t-1}$. The following lemma says that both the task-posterior $P_z(\cdot|h_{s,t})$ for any $z \in \mathcal{Z}$ and the meta-posterior $Q(\cdot|h_{1:s-1})$ depend only on the pulled arms according to π, but not the exact form of π.

Lemma 2. *Assume that there exists a σ-finite measure λ_ρ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that $\rho_{\theta,i}$ is absolutely continuous with respect to λ_ρ for all $\theta \in \Theta$ and $i \in [K]$. Then the task-posterior and meta-posterior exist and have the following form*

$$\begin{aligned} & P_z(S_1|h_{s,t}) = \frac{\int_{S_1} \prod_{j=1}^{t-1} p_{\theta_s, a_{s,j}}(x_{s,j})\,\mathrm{d}P_z(\theta_s)}{\int_{\Theta} \prod_{j=1}^{t-1} p_{\theta_s, a_{s,j}}(x_{s,j})\,\mathrm{d}P_z(\theta_s)}, \\ & Q(S_2|h_{1:s-1}) = \\ & \quad \frac{\int_{S_2} \left[\prod_{\ell=1}^{s-1} \prod_{j=1}^{n} \int_{\Theta} p_{\theta, a_{\ell,j}}(x_{\ell,j})\,\mathrm{d}P_z(\theta)\right] \mathrm{d}Q(z)}{\int_{\mathcal{Z}} \left[\prod_{\ell=1}^{s-1} \prod_{j=1}^{n} \int_{\Theta} p_{\theta, a_{\ell,j}}(x_{\ell,j})\,\mathrm{d}P_z(\theta)\right] \mathrm{d}Q(z)}, \end{aligned}$$

for any $S_1 \in \sigma(\Theta)$, $z \in \mathcal{Z}$, $S_2 \in \sigma(\mathcal{Z})$.

The proof is provided in Appendix C. `MetaTS` samples Z_s from the meta-posterior $Q_s = Q(\cdot|h_{1:s-1})$ at the beginning of each task $s \in [m]$ and then pulls arms according to the samples from $P_{Z_s}(\cdot|h_{s,t})$. The above lemma shows that to compute the posteriors we only need the distributions of rewards and integrate them over the environments (for the task-posterior) or over both the environments and the hidden variables (for the meta-posterior).These integrals can be derived analytically, for example, in the case of conjugate priors Sections 3.2 and 3.3.

4. Analysis

We bound the Bayes regret of `MetaTS` in Gaussian bandits (Section 3.3). This section is organized as follows. We state the bound and sketch its proof in Section 4.1, and discuss it in Section 4.2. In Section 4.3, we present the key lemmas. Finally, in Section 4.4, we discuss how our analysis can be extended beyond Gaussian bandits.

4.1. Regret Bound

We analyze `MetaTS` under the assumption that each arm is pulled at least once per task. Although this suffices to show benefits of meta-learning, it is conservative. A less conservative analysis would require understanding how Thompson sampling with a misspecified prior pulls arms. In particular, we would require a high-probability lower bound on the number of pulls of each arm. To the best of our knowledge, such as a bound does not exist and is non-trivial to derive. To guarantee that each arm is pulled at least once, we pull each arm in the last K rounds of each task. This is to avoid any interference with our posterior sampling analyses in the earlier rounds. Other than this, `MetaTS` is analyzed exactly as described in Sections 3.1 and 3.3.

Recall the following definitions in our setting (Section 3.3). The meta-prior is $Q = \mathcal{N}(\mathbf{0}, \sigma_q^2 I_K)$. The instance prior is $P_* = \mathcal{N}(\mu_*, \sigma_0^2 I_K)$, where $\mu_* \sim Q$ is chosen before the learning agent interact with the tasks. Then, in each task s, a problem instance is drawn i.i.d. as $\theta_{s,*} \sim P_*$. Our main result is the following Bayes regret bound.

Theorem 3. *The Bayes regret of* `MetaTS` *over m tasks with n rounds each is*

$$\begin{aligned} & R(m, n; P_*) \leq \\ & \quad c_1 \sqrt{K} \left(\sqrt{n + \sigma^2 \sigma_0^{-2} K} - \sqrt{\sigma^2 \sigma_0^{-2} K}\right) m + \\ & \quad c_2(\delta) c_3(\delta) K n^2 \sqrt{m} + \tilde{O}(Km + n) \end{aligned}$$

with probability at least $1 - (2m+1)\delta$, where

$$\begin{aligned} c_1 &= 4\sqrt{2\sigma^2 \log n}\,, \\ c_2(\delta) &= 2\left(\sqrt{2\sigma_q^2 \log(2K/\delta)} + \sqrt{2\sigma_0^2 \log n}\right), \\ c_3(\delta) &= 8\sqrt{(\sigma_0^2 + \sigma^2)\log(4K/\delta)/(\pi\sigma_0^2)}\,. \end{aligned}$$

The probability is over realizations of μ_*, $\theta_{s,*}$, and $\hat{\mu}_s$.

Proof. First, we bound the magnitude of μ_*. Specifically, since $\mu_* \sim \mathcal{N}(\mathbf{0}, \sigma_q^2 I_K)$, we have that

$$\|\mu_*\|_\infty \leq \sqrt{2\sigma_q^2 \log(2K/\delta)} \qquad (2)$$

holds with probability at least $1 - \delta$.

Now we fix task $s \geq 2$ and decompose its regret. Let $A_{s,*}$ be the optimal arm in instance $\theta_{s,*}$, $A_{s,t}$ be the pulled arm in round t by TS with misspecified prior P_s, and $\tilde{A}_{s,t}$ be the pulled arm in round t by TS with correct prior P_*. Then

$$\mathbb{E}\left[\sum_{t=1}^{n} \theta_{s,*}(A_{s,*}) - \theta_{s,*}(A_{s,t}) \,\middle|\, P_*\right] = R_{s,1} + R_{s,2}\,,$$

where

$$R_{s,1} = \mathbb{E}\left[\sum_{t=1}^{n} \theta_{s,*}(A_{s,*}) - \theta_{s,*}(\tilde{A}_{s,t}) \,\middle|\, P_*\right],$$
$$R_{s,2} = \mathbb{E}\left[\sum_{t=1}^{n} \theta_{s,*}(\tilde{A}_{s,t}) - \theta_{s,*}(A_{s,t}) \,\middle|\, P_*\right].$$

The term $R_{s,1}$ is the regret of hypothetical TS that knows P_*. This TS is introduced only for the purpose of analysis and is the optimal policy. The term $R_{s,2}$ is the difference in the expected n-round rewards of TS with priors P_s and P_*, and vanishes as the number of tasks s increases.

To bound $R_{s,1}$, we apply Lemma 4 with $\delta = 1/n$ and get

$$R_{s,1} \leq 4\sqrt{2\sigma^2 K \log n} \times \left(\sqrt{n + \sigma^2 \sigma_0^{-2} K} - \sqrt{\sigma^2 \sigma_0^{-2} K}\right) + \tilde{O}(K),$$

where $\tilde{O}(K)$ corresponds to the $c(\delta)$ term in Lemma 4. To bound $R_{s,2}$, we apply Lemma 5 with $\delta = 1/n$ and get

$$R_{s,2} \leq 2\left(\|\mu_*\|_\infty + \sqrt{2\sigma_0^2 \log n}\right)\sqrt{\frac{2}{\pi\sigma_0^2}} \times K n^2 \|\tilde{\mu}_s - \mu_*\|_\infty + \tilde{O}(K),$$

where $\tilde{O}(K)$ is the first term in Lemma 5, after we bound $\|\mu_*\|_\infty$ in it using (2).

Now we sum up our bounds on $R_{s,1} + R_{s,2}$ over all tasks $s \geq 2$ and get

$$c_1\sqrt{K}\left(\sqrt{n + \sigma^2\sigma_0^{-2}K} - \sqrt{\sigma^2\sigma_0^{-2}K}\right)m + c_2(\delta)\sqrt{\frac{2}{\pi\sigma_0^2}} K n^2 \sum_{s=2}^{m} \|\tilde{\mu}_s - \mu_*\|_\infty + \tilde{O}(Km),$$

where c_1 and $c_2(\delta)$ are defined in the main claim. Then we apply Lemma 6 to each term $\|\tilde{\mu}_s - \mu_*\|_\infty$ and have with probability at least $1 - m\delta$ that

$$\sum_{s=2}^{m} \|\tilde{\mu}_s - \mu_*\|_\infty \leq 4\sqrt{2(\sigma_0^2 + \sigma^2) m \log(4K/\delta)},$$

where $\sqrt{m}$ arises from summing up the $O(1/\sqrt{s})$ terms in Lemma 6, using Lemma 8 in Appendix B. This concludes the main part of the proof.

We finish with an upper bound on the regret in task 1 and the cost of pulling each arm once at the end of each task. This is can done as follows. Since $\theta_{s,*} \sim \mathcal{N}(\mu_*, \sigma_0^2 I_K)$,

$$\|\theta_{s,*} - \mu_*\|_\infty \leq \sqrt{2\sigma_0^2 \log(2K/\delta)} \tag{3}$$

holds with probability at least $1 - \delta$ in any task s. From (2) and (3), we have with a high probability that

$$\|\theta_{s,*}\|_\infty \leq \|\theta_{s,*} - \mu_*\|_\infty + \|\mu_*\|_\infty \leq 2\sqrt{(\sigma_q^2 + \sigma_0^2)\log(2K/\delta)}.$$

This yields a high-probability upper bound of

$$4\sqrt{(\sigma_q^2 + \sigma_0^2)\log(2K/\delta)}(n + Km)$$

on the regret in task 1 and pulling each arm once at the end of all tasks. This concludes our proof. □

4.2. Discussion

If we assume a "large m and n" regime, where the learning agent improves with more tasks but also needs to perform well in each task, the most important terms in Theorem 3 are those where m and n interact. Using these terms, our bound can be summarized as

$$\sqrt{K}\left[\sqrt{n + \sigma^2\sigma_0^{-2}K} - \sqrt{\sigma^2\sigma_0^{-2}K}\right]m + Kn^2\sqrt{m} \tag{4}$$

and holds with probability $1 - (2m + 1)\delta$ for any $\delta > 0$.

Our bound in (4) can be viewed as follows. The first term is the regret of Thompson sampling with the correct prior P_*. It is linear in the number of tasks m, since `MetaTS` solves m exploration problems. The second term captures the cost of learning P_*. Since it is sublinear in the number of tasks m, `MetaTS` is near optimal in the regime of "large m".

We compare our bound to two baselines. The first baseline is TS with a known prior P_*. The regret of this TS can be bounded using Lemma 4 and includes only the first term in (4). In the regime of "large m", this term dominates the regret of `MetaTS`, and thus `MetaTS` is near optimal.

The second baseline is *agnostic* Thompson sampling, which does use the structure $\theta_{s,*} \sim P_* \sim Q$. Instead, it marginalizes out Q. In our setting, this can be equivalently viewed as assuming $\theta_{s,*} \sim \mathcal{N}(\mathbf{0}, (\sigma_q^2 + \sigma_0^2)I_K)$. For this prior, the Bayes regret is $\mathbb{E}\left[R(m, n; P_*)\right]$, where the expectation is over $P_* \sim Q$. Again, we can apply Lemma 4 and show that $\mathbb{E}\left[R(m, n; P_*)\right]$ has an upper bound of

$$\sqrt{K}\left(\sqrt{n + \sigma^2\tilde{\sigma}^{-1}K} - \sqrt{\sigma^2\tilde{\sigma}^{-1}K}\right)m,$$

where $\tilde{\sigma}^2 = \sigma_q^2 + \sigma_0^2$. Clearly, $\tilde{\sigma}^2 > \sigma_0^2$; and therefore the difference of the above square roots is always larger than in (4). So, in the regime of "large m", `MetaTS` has a lower regret than this baseline.

4.3. Key Lemmas

Now we present the three key lemmas used in the proof of Theorem 3. They are proved in Appendix A.

The first lemma is a prior-dependent upper bound on the Bayes regret of TS.

Lemma 4. *Let θ_* be arm means in a K-armed Gaussian bandit that are generated as $\theta_* \sim P_* = \mathcal{N}(\mu_*, \sigma_0^2 I_K)$. Let A_* be the optimal arm under θ_* and A_t be the pulled arm in round t by TS with prior P_*. Then for any $\delta > 0$,*

$$\mathbb{E}\left[\sum_{t=1}^{n} \theta_*(A_*) - \theta_*(A_t)\right] \leq c(\delta) + 4\sqrt{2\sigma^2 K \log(1/\delta)}\left(\sqrt{n + \sigma^2 \sigma_0^{-2} K} - \sqrt{\sigma^2 \sigma_0^{-2} K}\right),$$

where $c(\delta) = 2\sqrt{2\sigma_0^2 \log(1/\delta)}K + \sqrt{2\sigma_0^2/\pi}Kn\delta$.

The effect of the prior is reflected in the difference of the square roots. As the prior width narrows and $\sigma_0 \to 0$, the difference decreases, which shows that a more concentrated prior leads to less exploration. The algebraic form of the bound is also expected. Roughly speaking, $\sqrt{\sigma^2 \sigma_0^{-2} K}$ is the sum of confidence interval widths in the Bayes regret analysis that cannot occur, because the prior width is σ_0.

The bound in Lemma 4 differs from other similar bounds in the literature (Lu & Van Roy, 2019). One difference is that the Cauchy-Schwarz inequality is not used in its proof. Therefore, σ_0 is in the square root instead of the logarithm. The resulting bound is tighter for $\sigma^2 \sigma_0^{-2} K \ll n$. Another difference is that information-theory arguments are not used in the proof. The dependence on σ_0 is a result of carefully characterizing the posterior variance of θ_* in each round. Our proof is simple and easy to follow.

The second lemma bounds the difference in the expected n-round rewards of TS with different priors.

Lemma 5. *Let θ_* be arm means in a K-armed Gaussian bandit that are generated as $\theta_* \sim P_* = \mathcal{N}(\mu_*, \sigma_0^2 I_K)$. Let $\mathcal{N}(\hat{\mu}, \sigma_0^2 I_K)$ and $\mathcal{N}(\tilde{\mu}, \sigma_0^2 I_K)$ be two TS priors such that $\|\hat{\mu} - \tilde{\mu}\|_\infty \leq \varepsilon$. Let $\hat{\theta}_t$ and $\tilde{\theta}_t$ be their respective posterior samples in round t, $\hat{A}_t$ and $\tilde{A}_t$ be the pulled arms under these samples. Then for any $\delta > 0$,*

$$\mathbb{E}\left[\sum_{t=1}^{n} \theta_*(\hat{A}_t) - \theta_*(\tilde{A}_t)\right] \leq 4\left(\sqrt{\frac{\sigma_0^2}{2\pi}} + \|\mu_*\|_\infty\right)Kn\delta + 2\left(\|\mu_*\|_\infty + \sqrt{2\sigma_0^2 \log(1/\delta)}\right)\sqrt{\frac{2}{\pi\sigma_0^2}}Kn^2\varepsilon\,.$$

The key dependence in Lemma 5 is that the bound is linear in the difference of prior means ε. The bound is also $O(n^2)$. Although this is unfortunate, it cannot be improved in general if we want to keep linear dependence on ε. The $O(n^2)$ dependence arises in the proof as follows. We bound the difference in the expected n-round rewards of TS with two different priors by the probability that the two TS instances deviate in each round multiplied by the maximum reward that can be earned from that round. The probability is $O(\varepsilon)$ and the maximum reward is $O(n)$. This bound is applied n times, in each round, and thus the $O(n^2\varepsilon)$ dependence.

The last lemma shows the concentration of meta-posterior sample means.

Lemma 6. *Let $\mu_* \sim \mathcal{N}(\mathbf{0}, \sigma_q^2 I_K)$ and the prior parameters in task s be sampled as $\tilde{\mu}_s \mid H_{1:s-1} \sim \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s)$. Then*

$$\|\tilde{\mu}_s - \mu_*\|_\infty \leq 2\sqrt{2\frac{\sigma_0^2 + \sigma^2}{(\sigma_0^2 + \sigma^2)\sigma_q^{-2} + s - 1}\log(4K/\delta)}$$

holds jointly over all tasks $s \in [m]$ with probability at least $1 - m\delta$.

The key dependence is that the bound is $O(1/\sqrt{s})$ in task s, which provides an upper bound on ε in Lemma 5. After we sum up these upper bounds over all $s \in [m]$ tasks, we get the $O(\sqrt{m})$ term in Theorem 3.

4.4. Beyond Gaussian Bandits

We analyze Gaussian bandits with a known prior covariance matrix (Section 4.1) because this simplifies algebra and is easy to interpret. We believe that a similar analysis can be conducted for other bandit problems based on the following high-level interpretation of our key lemmas (Section 4.3).

Lemma 4 says that more a concentrated prior in TS yields lower regret. This is expected in general, as less uncertainty about the problem instance leads to lower regret.

Lemma 5 says that the difference in the expected n-round rewards of TS with different priors can be bounded by the difference of the prior parameters. This is expected for any prior that is smooth in its parameters.

Lemma 6 says that the meta-posterior concentrates as the number of tasks increases. When each arm is pulled at least once per task, as we assume in Section 4.1, `MetaTS` gets at least one noisy observation of the prior per task, and any exponential-family meta-posterior would concentrate.

5. Experiments

We experiment with three problems. In each problem, we have $m = 20$ tasks with a horizon of $n = 200$ rounds. All results are averaged over 100 runs, where $P_* \sim Q$ in each run.

The first problem is a Bernoulli bandit with $K = 2$ arms and a categorical meta-prior (Section 3.2). We have $L = 2$

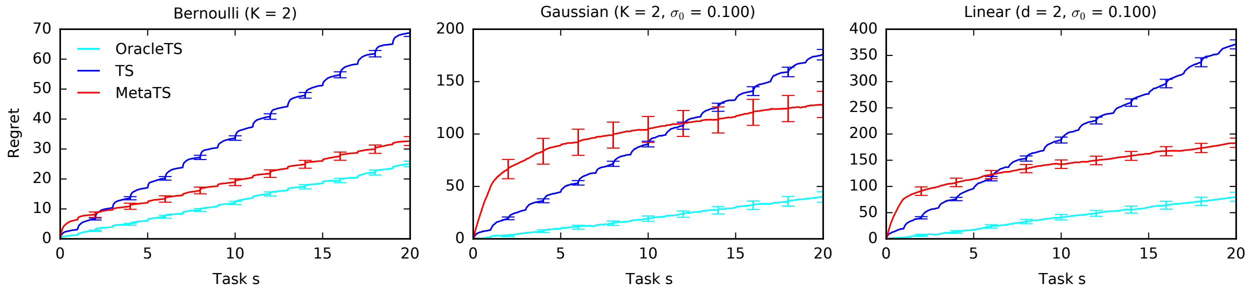

Figure 2. Comparison of `MetaTS` to two variants of Thompson sampling, where the instance prior P_* is known (`OracleTS`) and the meta-prior Q is marginalized out (`TS`).

instance priors, which are defined as

$$P^{(1)}(\theta) = \mathrm{Beta}(\theta_1; 6, 2)\,\mathrm{Beta}(\theta_2; 2, 6)\,,$$
$$P^{(2)}(\theta) = \mathrm{Beta}(\theta_1; 2, 6)\,\mathrm{Beta}(\theta_2; 6, 2)\,.$$

In instance prior $P^{(1)}$, arm 1 is more likely to be optimal than arm 2, while arm 1 is more likely to be optimal in prior $P^{(2)}$. The meta-prior is a categorical distribution $\mathrm{Cat}(w)$ where $w = (0.5, 0.5)$. This problem is designed such that if the agent knew P_*, it would know the optimal arm with high probability, and could significantly reduce exploration in future tasks.

The second problem is a Gaussian bandit with $K = 2$ arms and a Gaussian meta-prior (Section 3.3). The meta-prior width is $\sigma_q = 0.5$, the instance prior width is $\sigma_0 = 0.1$, and the reward noise is $\sigma = 1$. In this problem, $\sigma_q \gg \sigma_0$ and we expect major gains from meta-learning. In particular, based on our discussion in Section 4.2,

$$\begin{aligned} 5.85 &\approx \sqrt{n + \sigma^2\sigma_0^{-2}K} - \sqrt{\sigma^2\sigma_0^{-2}K} \\ &< \sqrt{n + \sigma^2(\sigma_0^2 + \sigma_q^2)^{-1}K} - \sqrt{\sigma^2(\sigma_0^2 + \sigma_q^2)^{-1}K} \\ &\approx 11.63\,. \end{aligned}$$

The third problem is a linear bandit in $d = 2$ dimensions with $K = 5d$ arms. We sample arm features uniformly at random from $[-0.5, 0.5]^d$. The meta-prior, prior, and noise are set as in the Gaussian experiment. The main difference is that $\theta_{s,*}$ is a parameter vector of a linear model, where the mean reward of arm x is $x^\top \theta_{s,*}$. The posterior updates are computed as described in Appendix D. Even in this more complex setting, they have a closed form.

We compare `MetaTS` to two baselines. The first, `OracleTS`, is idealized TS with the true prior P_*. This baseline shows the lowest attainable regret. The second baseline is agnostic TS, which does not use the structure of our problem. In the Gaussian and linear bandit experiments, we implement it as TS with prior $\mathcal{N}(\theta; \mathbf{0}, (\sigma_q^2 + \sigma_0^2)I_K)$, as this is a marginal distribution of $\theta_{s,*}$. In the Bernoulli bandit experiment, we use an uninformative prior $\prod_{i=1}^K \mathrm{Beta}(\theta_i; 1, 1)$, since the marginal distribution does not have a closed form. We call this baseline `TS`.

Our results are reported in Figure 2. We plot the cumulative regret as a function of the number of experienced tasks s, as it accumulates round-by-round within tasks. The regret of algorithms that do not learn μ_*, such as `OracleTS` and `TS`, is linear in s, since they solve s similar tasks using the same policy (Section 2). A lower slope of the regret indicates a better policy. Since `OracleTS` is optimal in our problems, no algorithm can have sublinear regret in s.

In all plots in Figure 2, we observe significant gains due to meta-learning P_*. `MetaTS` outperforms `TS`, which does not adapt to P_* and performs comparably to `OracleTS`. This can be seen from the slope of the regret. Specifically, the slope of the `MetaTS` regret approaches that of `OracleTS` after just a few tasks. The slopes of `TS` and `OracleTS` do not change, as these methods do not adapt between tasks.

In Appendix E, we report additional experimental results. We observe that the benefits of meta-learning are preserved as the number of arms K or dimensions d increases. However, as is expected, they diminish when the prior width σ_0 approaches the meta-prior width σ_q. In this case, there is little benefit from adapting to P_* and all methods perform similarly. We also experiment with misspecified `MetaTS` and show that the impact of the misspecification is relatively minor. This attests to the robustness of `MetaTS`.

6. Related Work

The closest related work is that of Bastani et al. (2019) who propose TS that learns an instance prior from a sequence of pricing experiments. Their approach is tailored to pricing and learns through forced exploration using a conservative variant of TS, resulting in a meta-learning algorithm that is more conservative and less general than our work. Bastani et al. (2019) also do not derive improved regret bounds due to meta-learning.

`MetaTS` is an instance of meta-learning (Thrun, 1996; 1998;

Baxter, 1998; 2000), where the agent learns to act under an unknown prior P_* from interactions with bandit instances. Earlier works on a similar topic are Azar et al. (2013) and Gentile et al. (2014), who proposed UCB algorithms for multi-task learning in the bandit setting. Multi-task learning in contextual bandits, where the arms are similar tasks, was studied by Deshmukh et al. (2017). Cella et al. (2020) proposed a LinUCB algorithm that meta-learns mean parameter vectors in linear models. Yang et al. (2020) studied a setting where the learning agent interacts with multiple bandit instances in parallel and tries to learn their shared subspace. A general template for sequential meta-learning is outlined in Ortega et al. (2019). Our work departs from most of the above approaches in two aspects. First, we have a posterior sampling algorithm that naturally represents the uncertainty in the unknown prior P_*. Second, we have a Bayes regret analysis. The shortcoming of the Bayes regret is that it is a weaker optimality criterion than the frequentist regret. To the best of our knowledge, this is the first work to propose meta-learning for Thompson sampling that is natural and has provable guarantees on improvement.

It is well known that the regret of bandit algorithms can be reduced by tuning (Vermorel & Mohri, 2005; Maes et al., 2012; Kuleshov & Precup, 2014; Hsu et al., 2019). All of these works are empirical and focus on the offline setting, where the bandit algorithms are optimized against a known instance distribution. Several recent approaches formulated learning of bandit policies as policy-gradient optimization (Duan et al., 2016; Boutilier et al., 2020; Kveton et al., 2020; Yang & Toni, 2020; Min et al., 2020). Notably, both Kveton et al. (2020) and Min et al. (2020) proposed policy-gradient optimization of TS. These works are in the offline setting and have no global optimality guarantees, except for some special cases (Boutilier et al., 2020; Kveton et al., 2020).

7. Conclusions

Thompson sampling (Thompson, 1933), a very popular and practical bandit algorithm (Chapelle & Li, 2012; Agrawal & Goyal, 2012; Russo et al., 2018), is parameterized by a prior, which is specified by the algorithm designer. We study a more general setting where the designer can specify an uncertain prior, and the actual prior is learned from sequential interactions with bandit instances. We propose MetaTS, a computationally-efficient algorithm for this problem. Our analysis of MetaTS shows the benefit of meta-learning and builds on a novel prior-dependent upper bound on the Bayes regret of Thompson sampling. MetaTS shows considerable promise in our synthetic experiments.

Our work is a step in the exciting direction of meta-learning state-of-the-art exploration algorithms with guarantees. It has several limitations that should be addressed in future work. First, our regret analysis relies only on a single pull of an arm per task. While this simplifies the analysis and is sufficient to show improvements due to meta-learning, it is inherently conservative. Second, our analysis is limited to Gaussian bandits and relies heavily on the properties of Gaussian posteriors. While we believe that a generalization is possible (Section 4.4), it is likely to be more algebraically demanding. Finally, we hope to analyze our method in contextual bandits. As we show in Appendix D, meta-posterior updates in linear bandits with Gaussian noise have a closed form. We believe that our work lays foundations for a potential analysis of this approach, which could yield powerful contextual bandit algorithms that adapt to an unknown problem class.

References

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

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

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

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

Baxter, J. Theoretical models of learning to learn. In *Learning to Learn*, pp. 71–94. Springer, 1998.

Baxter, J. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Boutilier, C., Hsu, C.-W., Kveton, B., Mladenov, M., Szepesvari, C., and Zaheer, M. Differentiable meta-learning of bandit policies. In *Advances in Neural Information Processing Systems 33*, 2020.

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

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

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

Duan, Y., Schulman, J., Chen, X., Bartlett, P., Sutskever, I., and Abbeel, P. RL^2: Fast reinforcement learning via slow reinforcement learning. *CoRR*, abs/1611.02779, 2016. URL http://arxiv.org/abs/1611.02779.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY, 2007.

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

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

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

Hsu, C.-W., Kveton, B., Meshi, O., Mladenov, M., and Szepesvari, C. Empirical Bayes regret minimization. *CoRR*, abs/1904.02664, 2019. URL http://arxiv.org/abs/1904.02664.

Kuleshov, V. and Precup, D. Algorithms for multi-armed bandit problems. *CoRR*, abs/1402.6028, 2014. URL http://arxiv.org/abs/1402.6028.

Kveton, B., Mladenov, M., Hsu, C.-W., Zaheer, M., Szepesvari, C., and Boutilier, C. Differentiable meta-learning in contextual bandits. *CoRR*, abs/2006.05094, 2020. URL http://arxiv.org/abs/2006.05094.

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

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

Lindley, D. and Smith, A. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(1):1–41, 1972.

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

Maes, F., Wehenkel, L., and Ernst, D. Meta-learning of exploration/exploitation strategies: The multi-armed bandit case. In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence*, pp. 100–115, 2012.

Min, S., Moallemi, C., and Russo, D. Policy gradient optimization of Thompson sampling policies. *CoRR*, abs/2006.16507, 2020. URL http://arxiv.org/abs/2006.16507.

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

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

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

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

Thrun, S. *Explanation-Based Neural Network Learning - A Lifelong Learning Approach*. PhD thesis, University of Bonn, 1996.

Thrun, S. Lifelong learning algorithms. In *Learning to Learn*, pp. 181–209. Springer, 1998.

Tulcea, C. I. Mesures dans les espaces produits. *Atti Acad. Naz. Lincei Rend. Cl Sci. Fis. Mat. Nat*, 8(7), 1949.

Vermorel, J. and Mohri, M. Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of the 16th European Conference on Machine Learning*, pp. 437–448, 2005.

Yang, J., Hu, W., Lee, J., and Du, S. Provable benefits of representation learning in linear bandits. *CoRR*, abs/2010.06531, 2020. URL http://arxiv.org/abs/2010.06531.

Yang, K. and Toni, L. Differentiable linear bandit algorithm. *CoRR*, abs/2006.03000, 2020. URL http://arxiv.org/abs/2006.03000.