Meta-Learning for Simple Regret Minimization

Javad Azizi[1], **Branislav Kveton**[2], **Mohammad Ghavamzadeh**[3], **Sumeet Katariya**[2]

[1]University of Southern California
[2]Amazon
[3]Google Research
azizim@usc.edu, bkveton@amazon.com, ghavamza@google.com, katsumee@amazon.com

Abstract

We develop a meta-learning framework for simple regret minimization in bandits. In this framework, a learning agent interacts with a sequence of bandit tasks, which are sampled i.i.d. from an unknown prior distribution, and learns its meta-parameters to perform better on future tasks. We propose the first Bayesian and frequentist meta-learning algorithms for this setting. The Bayesian algorithm has access to a prior distribution over the meta-parameters and its meta simple regret over m bandit tasks with horizon n is mere $\tilde{O}(m/\sqrt{n})$. On the other hand, the meta simple regret of the frequentist algorithm is $\tilde{O}(\sqrt{m}n + m/\sqrt{n})$. While its regret is worse, the frequentist algorithm is more general because it does not need a prior distribution over the meta-parameters. It can also be analyzed in more settings. We instantiate our algorithms for several classes of bandit problems. Our algorithms are general and we complement our theory by evaluating them empirically in several environments.

1 Introduction

We study the problem of *simple regret minimization (SRM)* in a *fixed-horizon (budget) setting* (Audibert and Bubeck 2010; Kaufmann, Cappé, and Garivier 2016). The learning agent interacts sequentially with m such tasks, where each task has a horizon of n rounds. The tasks are sampled i.i.d. from a prior distribution P_*, which makes them similar. We study a meta-learning (Thrun 1996, 1998; Baxter 1998, 2000) variant of the problem, where the prior distribution P_* is unknown, and the learning agent aims to learn it to reduce its regret on future tasks.

This problem is motivated by practical applications, such as online advertising, recommender systems, hyper-parameter tuning, and drug repurposing (Hoffman, Shahriari, and Freitas 2014; Mason et al. 2020; Réda, Kaufmann, and Delahaye-Duriez 2021; Alieva, Cutkosky, and Das 2021), where bandit models are popular due to their simplicity and efficient algorithms. These applications include a test phase separated from the commercialization phase, and one aims at minimizing the regret of the commercialized product (simple regret) rather than the cumulative regret in the test phase (Audibert and Bubeck 2010). In all of these, the exploration phase is limited by a fixed horizon: the budget for estimating click rates on ads is limited, or a hyper-parameter tuning task has only a limited amount of resources (Alieva, Cutkosky, and Das 2021). Meta-learning can result in more efficient exploration when the learning agent solves similar tasks over time.

To understand the benefits of meta-learning, consider the following example. Repeated A/B tests are conducted on a website to improve customer engagement. Suppose that the designers always propose a variety of website designs to test. However, dark designs tend to perform better than light ones, and thus a lot of customer traffic is repeatedly wasted to discover the same pattern. One solution to reducing waste is that the designers to stop proposing light designs. However, these designs are sometimes better. A more principled solution is to automatically adapt the prior P_* in A/B tests to promote dark designs unless proved otherwise by evidence. This is the key idea in the proposed solution in this work.

We make the following contributions. First, we propose a general meta-learning framework for fixed-horizon SRM in Section 2. While several recent papers studied this problem in the cumulative regret setting (Bastani, Simchi-Levi, and Zhu 2019; Cella, Lazaric, and Pontil 2020; Kveton et al. 2021; Basu et al. 2021; Simchowitz et al. 2021), this work is the first application of meta-learning to SRM. We develop general Bayesian and frequentist algorithms for this problem in Sections 3 and 4. Second, we show that our Bayesian algorithm, which has access to a prior over the meta-parameters of P_*, has meta simple regret $\tilde{O}(m/\sqrt{n})$ over m bandit tasks with horizon n. Our frequentist algorithm is more general because it does not need a prior distribution over the meta-parameters. However, we show that its meta simple regret is $\tilde{O}(\sqrt{m}n+m/\sqrt{n})$, and thus, worse than that of the Bayesian algorithm. In Section 4.2, we present a lower bound showing that this is unimprovable in general. Third, we instantiate both algorithms in multi-armed and linear bandits in Section 5. These instances highlight the trade-offs of the Bayesian and frequentist approaches, a provably lower regret versus more generality. Finally, we complement our theory with experiments (Section 7), which show the benefits of meta-learning and confirm that the Bayesian approaches are superior whenever implementable.

Some of our contributions are of independent interest. For instance, our analysis of the meta SRM algorithms is based on a general reduction from cumulative regret minimization in

Section 3.1, which yields novel and easily implementable algorithms for Bayesian and frequentist SRM, based on *Thompson sampling (TS)* and *upper confidence bounds (UCBs)* (Lu and Van Roy 2019). To the best of our knowledge, only Komiyama et al. (2021) studied Bayesian SRM before (Section 6). In Section 5.2, we also extend the analysis of frequentist meta-learning in Simchowitz et al. (2021) to structured bandit problems.

2 Problem Setup

In meta SRM, we consider m bandit problems with arm set $\mathcal{A}$ that appear sequentially and each is played for n rounds. At the beginning of each task (bandit problem) $s \in [m]$, the mean rewards of its arms $\mu_s \in \mathbb{R}^{\mathcal{A}}$ are sampled i.i.d. from a prior distribution P_*. We define $[m] = \{1, 2, \cdots, m\}$ for any integer m. We apply a base SRM algorithm, `alg`, to task s and denote this instance by $\texttt{alg}_s$. The algorithm interacts with task s for n rounds. In round $t \in [n]$ of task s, $\texttt{alg}_s$ pulls an arm $A_{s,t} \in \mathcal{A}$ and observes its reward $Y_{s,t}(A_{s,t})$, where $\mathbb{E}[Y_{s,t}(a)] = \mu_s(a)$. We assume that $Y_{s,t}(a) \sim \nu(a; \mu_s)$ where $\nu(\cdot; \mu_s)$ is the reward distribution of all arms with parameter (mean) μ_s. After the n rounds the algorithm returns arm $\hat{A}_{\texttt{alg}_s}$ or simply $\hat{A}_s$ as the *best arm*. Let $A^*_s = \arg\max_{a \in \mathcal{A}} \mu_s(a)$ be the best arm in task s. We define the *per-task simple regret* for task s as

$$\mathrm{SR}_s(n, P_*) = \mathbb{E}_{\mu_s \sim P_*} \mathbb{E}_{\mu_s}[\Delta_s], \tag{1}$$

where $\Delta_s = \mu_s(A^*_s) - \mu_s(\hat{A}_s)$. The outer expectation is w.r.t. the randomness of the task instance, and the inner one is w.r.t. the randomness of rewards and algorithm. This is the common frequentist simple regret averaged over instances drawn from P_*.

In the *frequentist* setting, we assume that P_* is unknown but fixed, and define the *frequentist meta simple regret* as

$$\mathrm{SR}(m, n, P_*) = \sum_{s=1}^{m} \mathrm{SR}_s(n, P_*) \ . \tag{2}$$

In the *Bayesian* setting, we still assume that P_* is unknown. However, we know that it is sampled from a known *meta prior* Q. We define *Bayesian meta simple regret* as

$$\mathrm{BSR}(m, n) = \mathbb{E}_{P_* \sim Q}[\mathrm{SR}(m, n, P_*)] \ . \tag{3}$$

3 Bayesian Meta-SRM

In this section, we present our Bayesian meta SRM algorithm (`B-metaSRM`), whose pseudo-code is in Algorithm 1. The key idea is to deploy `alg` for each task with an adaptively refined prior learned from the past interactions, which we call an *uncertainty-adjusted prior*, $P_s(\mu)$. This is an approximation to P_* and it is the posterior density of μ_s given the history up to task s. At the beginning of task s, `B-metaSRM` instantiates `alg` with P_s, denoted as $\texttt{alg}_s = \texttt{alg}(P_s)$, and uses it to solve task s.

The base algorithm `alg` is *Thompson Sampling (TS)* or *Bayesian UCB (BayesUCB)* (Lu and Van Roy 2019). During its execution, $\texttt{alg}_s$ keeps updating its posterior over μ_s as $P_{s,t}(\mu_s) \propto \mathcal{L}_{s,t}(\mu_s) P_s(\mu_s)$, where $\mathcal{L}_{s,t}(\mu_s) = \prod_{\ell=1}^{t} \mathbf{P}(Y_{s,\ell} | A_{s,\ell}, \mu_s)$ is the likelihood of observations in task s up to round t under task parameter μ_s. TS pulls the arms proportionally to being the best w.r.t. the posterior. More precisely, it samples $\tilde{\mu}_{s,t} \sim P_{s,t}$ and then pulls arm $A_{s,t} \in \arg\max_{a \in \mathcal{A}} \tilde{\mu}_{s,t}(a)$. BayesUCB is the same but it pulls the arm with largest Bayesian upper confidence bound (see Appendix C and Eq. (12) for details).

The critical step is how P_s is updated. Let θ_* be the parameter of P_* . At task s, `B-metaSRM` maintains a posterior density over the parameter θ_*, called *meta-posterior* $Q_s(\theta)$, and uses it to compute $P_s(\mu)$. We use the following recursive rule from Proposition 1 of Basu et al. (2021) to update Q_s and P_s.

Proposition 1. *Let $\mathcal{L}_{s-1}(\cdot) = \mathcal{L}_{s-1,n}(\cdot)$ be the likelihood of observations right before the start of task s. We let P_θ be the prior distribution parameterized by θ. Then* `alg` *computes Q_s and P_s as*

$$Q_s(\theta) = \int_\mu \mathcal{L}_{s-1}(\mu) P_\theta(\mu) \mathrm{d}\kappa_2(\mu) Q_{s-1}(\theta), \qquad \forall \theta \tag{4}$$

$$P_s(\mu) = \int_\theta P_\theta(\mu) Q_s(\theta) \mathrm{d}\kappa_1(\theta), \qquad \forall \mu \tag{5}$$

where κ_1 and κ_2 are the probability measures of θ and μ. We initialize Eq. (4) *with $\mathcal{L}_0 = 1$ and $Q_0 = Q$, where Q is the meta prior.*

Note that this update rule is computationally efficient for Gaussian prior with Gaussian meta-prior, but not many other distributions. This computational issue can limit the applicability of our Bayesian algorithm.

When task s ends, $\texttt{alg}_s$ returns the best arm $\hat{A}_{\texttt{alg}_s}$ by sampling from the distribution

$$\hat{A}_{\texttt{alg}_s} \sim \rho_s, \qquad \rho_s(a) := \frac{N_{a,s}}{n}, \tag{6}$$

where $N_{a,s} := |\{t \in [n] : A_{s,t} = a\}|$ is the number of rounds where arm a is pulled. That is, the algorithm chooses the arms proportionally to their number of pulls. This decision rule facilitates the analysis of our algorithms based on a reduction from cumulative to simple regret. We develop this reduction in Section 3.1 and show that per-task simple regret is essentially the cumulative regret divided by n. This yields novel algorithms for Bayesian and frequentist SRM with guarantees.

3.1 Cumulative to Simple Regret Reduction

Fix task s and consider an algorithm that pulls a sequence of arms $(A_{s,t})_{t \in [n]}$. Let its per-task cumulative regret with prior P be $R_s(n, P) := \mathbb{E}_{\mu_s \sim P} \mathbb{E}_{\mu_s}\left[n\mu_s(A^*_s) - \sum_{t=1}^{n} \mu_s(A_{s,t})\right]$, where the inner expectation is taken over the randomness in the rewards and algorithm. Now suppose that at the end of the task, we choose arm a with probability $\rho_s(a)$ and declare it to be the best arm $\hat{A}_s$. Then the per-task simple regret of this procedure is bounded as follows.

Proposition 2 (Cumulative to Simple Regret)**.** *For task s with n rounds, if we return an arm with probability proportional to its number of pulls as the best arm, the per-task simple regret with prior P is* $\mathrm{SR}_s(n, P) = R_s(n, P)/n$.

Algorithm 1: Bayesian Meta-SRM (B-metaSRM)

Input: Meta prior Q, base algorithm alg
Initialize: Meta posterior $Q_0 \leftarrow Q$
for $s = 1, \dots, m$ **do**
  Receive the current task s, $\mu_s \sim P_*$
  Compute meta posterior Q_s using Eq. (4)
  Compute uncertainty-adjusted prior P_s using Eq. (5)
  Instantiate alg for task s, $\texttt{alg}_s \leftarrow \texttt{alg}(P_s)$
  Run $\texttt{alg}_s$ for n rounds
  Return the best arm $\hat{A}_{\texttt{alg}_s} \sim \rho_s$ using Eq. (6)
end for

We prove this proposition in Appendix B using the linearity of expectation and properties of ρ_s. Note that Proposition 2 applies to both frequentist and Bayesian *meta* simple regret. This is because the former is a summation of SR_s over tasks, and the latter is achieved by taking an expectation of the former over P_*.

3.2 Bayesian Regret Analysis

Our analysis of B-metaSRM is based on results in Basu et al. (2021) and Lu and Van Roy (2019), combined with Section 3.1. Specifically, let $\Gamma_{s,t}$ be an information-theoretic constant independent of m and n that bounds the instant regret of the algorithm at round t of task s. We defer its precise definition to Appendix C as it is only used in the proofs. The following generic bound for the Bayesian meta simple regret of B-metaSRM holds.

Theorem 3 (Information Theoretic Bayesian Bound)**.** *Let $\{\Gamma_s\}_{s\in[m]}$ and Γ be non-negative constants, such that $\Gamma_{s,t} \leq \Gamma_s \leq \Gamma$ holds for all $s \in [m]$ and $t \in [n]$ almost surely. Then, the Bayesian meta simple regret (Eq. 3) of* B-metaSRM *satisfies*

$$\mathrm{BSR}(m,n) \leq \Gamma\sqrt{\frac{m}{n}\,\mathrm{I}(\theta_*;\tau_{1:m})} + \sum_{s=1}^{m}\Gamma_s\sqrt{\frac{\mathrm{I}(\mu_s;\tau_s|\theta_*,\tau_{1:s-1})}{n}} + \sum_{s=1}^{m}\sum_{t=1}^{n}\frac{\mathbb{E}[\beta_{s,t}]}{n}, \tag{7}$$

where $\tau_{1:s} = \oplus_{\ell=1}^{s}(A_{\ell,1}, Y_{\ell,1}, \cdots, A_{\ell,n}, Y_{\ell,n})$ is the trajectory up to task s, τ_s is similarly defined for the history only in task s, and $\mathrm{I}(\cdot;\cdot)$ and $\mathrm{I}(\cdot;\cdot|\cdot)$ are mutual information and conditional mutual information, respectively.

The proof is in Appendix C. It builds on the analysis in Basu et al. (2021) and uses our reduction in Section 3.1. Our reduction readily applies to Bayesian meta simple regret by linearity of expectation.

The first term in Eq. (7) is the price for learning the prior parameter θ_* and the second one is the price for learning the mean rewards of tasks $(\mu_s)_{s\in[m]}$ given known θ_*. It has been shown in many settings that the mutual information terms grow slowly with m and n (Lu and Van Roy 2019; Basu et al. 2021), and thus the first term is $\tilde{O}(\sqrt{m/n})$ and negligible. The second term is $\tilde{O}(m/\sqrt{n})$, since we solve m independent problems, each with $\tilde{O}(1/\sqrt{n})$ simple regret. In Section 5.2, we discuss a bandit environment where $\Gamma_{s,t}$ and $\beta_{s,t}$ are such that the last term of the bound is comparable to the rest. This holds in several other environments discussed in Lu and Van Roy (2019); Basu et al. (2021), and Liu et al. (2022).

Algorithm 2: Frequentist Meta-SRM (f-metaSRM)

Input: Exploration strategy explore, base algorithm alg
Initialize: $\tilde{\tau}_1 \leftarrow \emptyset$
for $s = 1, \dots, m$ **do**
  Receive the current task s, $\mu_s \sim P_*$
  Explore the arms using explore
  Append explored arms and their observations to $\tilde{\tau}_s$
  Compute $\hat{\theta}_s$ using $\tilde{\tau}_s$ as an estimate of θ_*
  Instantiate alg for task s, $\texttt{alg}_s \leftarrow \texttt{alg}(\hat{\theta}_s)$
  Run $\texttt{alg}_s$ for the rest of the n rounds
  Return the best arm $\hat{A}_{\texttt{alg}_s} \sim \rho_s$ using Eq. (6)
  $\tilde{\tau}_{s+1} \leftarrow \tilde{\tau}_s$
end for

4 Frequentist Meta-SRM

In this section, we present our frequentist meta SRM algorithm (f-metaSRM), whose pseudo-code is in Algorithm 2. Similarly to B-metaSRM, f-metaSRM uses TS or UCB as its base algorithm alg. However, it directly estimates its prior parameter, instead of maintaining a meta-posterior. At the beginning of task $s \in [m]$, f-metaSRM explores the arms for a number of rounds using an *exploration strategy* denoted as explore. This strategy depends on the problem class and we specify it for two classes in Section 5. f-metaSRM uses samples collected in the exploration phase of all the tasks up to task s, $\tilde{\tau}_s$, to update its estimate of the prior parameter $\hat{\theta}_s$. Then, it instantiates the base algorithm with this estimate, denoted as $\texttt{alg}_s$ =alg $(\hat{\theta}_s)$, and uses $\texttt{alg}_s$ for the rest of the rounds of task s. Here $\texttt{alg}(\theta) := \texttt{alg}(P_\theta)$ is the base algorithm alg instantiated with prior parameter θ (Note that we used a slightly different parameterization of alg compared to Section 3). When task s ends, $\texttt{alg}_s$ returns the best arm $\hat{A}_{\texttt{alg}_s}$ by sampling from the probability distribution ρ_s defined in Eq. (6).

While B-metaSRM uses a Bayesian posterior to maintain its estimate of θ_*, f-metaSRM relies on a frequentist approach. Therefore, it applies to settings where computing the posterior is not computationally feasible. Moreover, we can analyze f-metaSRM for general settings beyond Gaussian bandits.

4.1 Frequentist Regret Analysis

In this section, we prove an upper bound for the frequentist meta simple regret (Eq. 2) of f-metaSRM with TS alg. To start, we bound the per-task simple regret of alg relative to *oracle* that knows θ_*. To be more precise, this is the difference between the means of arms returned by alg instantiated with some prior parameter θ and the true prior parameter θ_*.

The *total variation* (TV) distance for two distributions P and P' over the same probability space $(\Omega, \mathcal{F})$[1] is defined as

[1] Ω is the sample space and $\mathcal{F}$ is the sigma-algebra.

$\mathrm{TV}(P \mid\mid P') := \sup_{E \in \mathcal{F}} |P(E) - P'(E)|$. We use TV to measure the distance between the estimated and true priors. We fix task s and drop subindexing by s. In the following, we bound the per-task simple regret of $\texttt{alg}(\theta)$ relative to oracle $\texttt{alg}(\theta_*)$.

Theorem 4. *Suppose P_{θ_*} is the true prior of the tasks and satisfies $P_{\theta_*}(\mathrm{diam}(\mu) \leq B) = 1$, where $\mathrm{diam}(\mu) := \sup_{a \in \mathcal{A}} \mu(a) - \inf_{a \in \mathcal{A}} \mu(a)$. Let θ be a prior parameter, such that $\mathrm{TV}(P_{\theta_*} \mid\mid P_\theta) = \epsilon$. Also, let $\hat{A}_{\texttt{alg}(\theta_*)}$ and $\hat{A}_{\texttt{alg}(\theta)}$ be the arms returned by $\texttt{alg}(\theta_*)$ and $\texttt{alg}(\theta)$, respectively. Then we have*

$$\mathbb{E}_{\mu \sim P_{\theta_*}} \mathbb{E}\big[\mu(\hat{A}_{\texttt{alg}(\theta_*)}) - \mu(\hat{A}_{\texttt{alg}(\theta)})\big] \leq 2n\epsilon B. \quad (8)$$

Moreover, if the prior is coordinate-wise σ_0^2-sub-Gaussian (Definition 14 in Appendix E), then we may write the RHS of Eq. (8) as $2n\epsilon\Big(\mathrm{diam}\big(\mathbb{E}_{\theta_*}[\mu]\big) + \sigma_0\Big(8 + 5\sqrt{\log \frac{|\mathcal{A}|}{\min(1, 2n\epsilon)}}\Big)\Big)$, where $\mathbb{E}_{\theta_*}[\mu]$ is the expectation of the mean reward of the arms, μ, given the true prior θ_*.

The proof in Appendix E uses the fact that TS is a 1-Monte Carlo algorithm, as defined by Simchowitz et al. (2021). It builds on Simchowitz et al. (2021) analysis of the cumulative regret, and extends it to simple regret. We again use our reduction in Section 3.1, which shows how it can be applied to a frequentist setting.

Theorem 4 shows that an ϵ prior misspecification leads to $O(n\epsilon)$ simple regret cost in `f-metaSRM`. The constant terms in the bounds depend on the prior distribution. In particular, for a bounded prior, they reflect the variability (diameter) of the expected mean reward of the arms. Moreover, under a sub-Gaussian prior, the bound depends logarithmically on the number of arms $|\mathcal{A}|$ and sub-linearly on the prior variance proxy σ_0^2.

Next, we bound the frequentist meta simple regret (Eq. 2) of `f-metaSRM`.

Corollary 4.1 (Meta Simple Regret of `f-metaSRM`)**.** *Let the* `explore` *strategy in Algorithm 2 be such that $\epsilon_s = \mathrm{TV}(P_{\theta_*} \mid\mid P_{\hat{\theta}_s}) = O(1/\sqrt{s})$ for each task $s \in [m]$. Then the frequentist meta simple regret of* `f-metaSRM` *is bounded as*

$$\mathrm{SR}(m, n, P_{\theta_*}) = O\left(2\sqrt{m}nB + m\sqrt{|\mathcal{A}|/n}\right). \quad (9)$$

The proof is in Appendix E and decomposes the frequentist meta simple regret into two terms: (i) the per-task simple regret of $\texttt{alg}(\hat{\theta}_s)$ relative to oracle $\texttt{alg}(\theta_*)$ in task s, which we bound in Theorem 4, and (ii) the meta simple regret of the oracle $\texttt{alg}(\theta_*)$, which we bound using our cumulative regret to simple regret reduction (Section 3.1).

The $O(\sqrt{mn})$ term is the price of estimating the prior parameter, because it is the per-task simple regret relative to the oracle. The $O(m\sqrt{|\mathcal{A}|/n})$ term is the meta simple regret of the oracle over m tasks.

Comparing to our bound in Theorem 3, `B-metaSRM` has a lower regret of $O(\sqrt{m/n} + m/\sqrt{n}) = O(m/\sqrt{n})$. More precisely, only the price for learning the prior is different as both bounds have $O(m/\sqrt{n})$ terms. Note that despite its smaller regret bound, `B-metaSRM` may not be computationally feasible for arbitrary distributions and priors, while `f-metaSRM` is since it directly estimates the prior parameter using frequentist techniques.

4.2 Lower Bound

In this section, we prove a lower bound on the relative per-task simple regret of a γ-shot TS algorithm, i.e., a TS algorithm that takes $\gamma \in \mathbb{N}$ samples (instead of 1) from the posterior in each round. This lower bound compliments our upper bound in Theorem 4 and shows that Eq. (8) is near-optimal. The proof of our lower bound builds on a cumulative regret lower bound in Theorem 3.3 of Simchowitz et al. (2021) and extends it to simple regret. We present the proof in Appendix E.2.

Theorem 5 (Lower Bound)**.** *Let $\texttt{TS}_\gamma(\theta)$ be a γ-shot TS algorithm instantiated with the prior parameter θ. Also let P_θ and $P_{\theta'}$ be two task priors. Let $\mu \in [0,1]^{\mathcal{A}}$ and fix a tolerance $\eta \in (0, \frac{1}{4})$. Then there exists a universal constant c_0 such that for any horizon $n \geq \frac{c_0}{\eta}$, number of arms $|\mathcal{A}| = n\lceil \frac{c_0}{\eta} \rceil$, and error $\epsilon \leq \frac{\eta}{c_0 \gamma n}$, we have $\mathrm{TV}(P_\theta \mid\mid P_{\theta'}) = \epsilon$ and the difference of per-task simple regret of $\texttt{TS}_\gamma(\theta)$ and $\texttt{TS}_\gamma(\theta')$ satisfies $\mathbb{E}[\mu(\hat{A}_{\texttt{TS}_\gamma(\theta)})] - \mathbb{E}[\mu(\hat{A}_{\texttt{TS}_\gamma(\theta')})] \geq (\frac{1}{2} - \eta)\gamma n\epsilon$.*

This lower bound holds for any setting with large enough n and $|\mathcal{A}| = O(n^2)$, and a small prior misspecification error $\epsilon = O(1/n^2)$. This makes it relatively general.

5 Meta-Learning Examples

In this section, we apply our algorithms to specific priors and reward distributions. The main two are the Bernoulli and linear (contextual) Gaussian bandits. We analyze `f-metaSRM` in an *explore-then-commit* fashion, where `f-metaSRM` estimates the prior using `explore` in the first m_0 tasks and then commits to it. This is without loss of generality and only for simplicity.

5.1 Bernoulli Bandits

We start with a Bernoulli multi-armed bandit (MAB) problem, as TS was first analyzed in this setting (Agrawal and Goyal 2012). Consider Bernoulli rewards with beta priors for $\mathcal{A} = [K]$ arms. In particular, assume that the prior is $P_* = \bigotimes_{a \in \mathcal{A}} \mathrm{Beta}(\alpha_a^*, \beta_a^*)$. Therefore, α_a^* and β_a^* are the prior parameters of arm a and the arm mean $\mu_s(a)$ is the probability of getting reward 1 for arm a when it is pulled. $\mathrm{Beta}(\alpha, \beta)$ is the beta distribution with a support on $(0, 1)$ with parameters $\alpha > 0$ and $\beta > 0$.

`B-metaSRM` in this setting does not have a computationally tractable meta-prior (Basu et al. 2021). We can address this in practice by discretization and using TS as described in Section 3.4 of Basu et al. (2021). However, the theoretical analysis for this case does not exist. This is because a computationally tractable prior for a product of beta distributions does not exist. It is challenging to generalize our Bayesian approach to this class of distributions as we require more than the standard notion of conjugacy.

In the contrary, `f-metaSRM` directly estimates the beta prior parameters, $(\alpha_a^*)_{a \in \mathcal{A}}$ and $(\beta_a^*)_{a \in \mathcal{A}}$ based on the observed Bernoulli rewards as follows. The algorithm explores only in $m_0 \leq m$ tasks. `explore` samples arm 1 in the

first t_0 rounds of first m_0/K tasks, and arm 2 in the next m_0/K tasks similarly, and so on for arm 3 to K. In other words, `explore` samples arm $a \in [K]$ in the first t_0 rounds of a'th batch of size m_0/K tasks. Let X_s denote the cumulative reward collected in the first t_0 rounds of task s. Then, the random variables $X_1, \ldots, X_{m_0/K}$ are i.i.d. draws from a Beta-Binomial distribution (BBD) with parameters $(\alpha_1^*, \beta_1^*, t_0)$, where t_0 denotes the number of trials of the binomial component. Similarly, $X_{(m_0/K)+1}, \cdots, X_{2m_0/K}$ are i.i.d. draws from a BBD with parameters $(\alpha_2^*, \beta_2^*, t_0)$. In general, $X_{(a-1)(m_0/K)+1}, \cdots, X_{am_0/K}$ are i.i.d. draws from a BBD with parameters $(\alpha_a^*, \beta_a^*, t_0)$. Knowing this, it is easy to calculate the prior parameters for each arm using the method of moments (Tripathi, Gupta, and Gurland 1994). The detailed calculations are in Appendix D. We prove the following result in Appendix E.3.

Corollary 5.1 (Frequentist Meta Simple Regret, Bernoulli)**.** *Let* `alg` *be a TS algorithm that uses the method of moments described and detailed in Appendix D, to estimate the prior parameters with* $m_0 \geq \frac{C|\mathcal{A}|^2 \log(|\mathcal{A}|/\delta)}{\epsilon^2}$ *exploration tasks (explore-then-commit). Then the frequentist meta simple regret of* `f-metaSRM` *satisfies* $\mathrm{SR}(m, n, P_{\theta_*}) = O\big(2mn\epsilon + m\sqrt{\frac{|\mathcal{A}|\log(n)}{n}} + m_0\big)$*, for* $m \geq m_0$ *with probability at least* $1 - \delta$*.*

With small enough ϵ, the bound shows $\tilde{O}(m/\sqrt{|\mathcal{A}|/n})$ scaling which we conjecture is the best an oracle that knows the correct prior of each task could do in expectation. The bound seems to be only sublinear in n if $\epsilon = O(1/n^{3/2})$. However, since $\epsilon \propto m_0^{-1/2}$ and we know $\sum_{z=1}^{m} z^{-1/2} = m^{1/2}$, if the exploration continues in all tasks, the regret bound above simplifies to $O\Big(\sqrt{m}n + m\sqrt{\frac{|\mathcal{A}|\log(n)}{n}}\Big)$.

5.2 Linear Gaussian Bandits

In this section, we consider linear contextual bandits. Suppose that each arm $a \in \mathcal{A}$ is a vector in $\mathbb{R}^d$ and $|\mathcal{A}| = K$. Also, assume $\nu_s(a; \mu_s) = \mathcal{N}(a^\top \mu_s, \sigma^2)$, i.e., with a little abuse of notation $\mu_s(a) = a^\top \mu_s$, where μ_s is the parameter of our linear model. A conjugate prior for this problem class is $P_* = \mathcal{N}(\theta_*, \Sigma_0)$, where $\Sigma_0 \in \mathbb{R}^{d\times d}$ is known and we learn $\theta_* \in \mathbb{R}^d$.

In the Bayesian setting, we assume that the meta-prior is $Q = \mathcal{N}(\psi_q, \Sigma_q)$, where $\psi_q \in \mathbb{R}^d$ and $\Sigma_q \in \mathbb{R}^{d\times d}$ are both known. In this case, the meta-posterior is $Q_s = \mathcal{N}(\hat{\theta}_s, \hat{\Sigma}_s)$, where $\hat{\theta}_s \in \mathbb{R}^d$ and $\hat{\Sigma}_s \in \mathbb{R}^{d\times d}$ are calculated as

$$\hat{\theta}_s = \hat{\Sigma}_s\Big(\Sigma_q^{-1}\psi_q + \sum_{\ell=1}^{s-1}\frac{B_\ell}{\sigma^2} - \frac{V_\ell}{\sigma^2}\Big(\Sigma_0^{-1} + \frac{V_\ell}{\sigma^2}\Big)^{-1}\frac{B_\ell}{\sigma^2}\Big),$$

$$\hat{\Sigma}_s^{-1} = \Sigma_q^{-1} + \sum_{\ell=1}^{s-1}\frac{V_\ell}{\sigma^2} - \frac{V_\ell}{\sigma^2}\Big(\Sigma_0^{-1} + \frac{V_\ell}{\sigma^2}\Big)^{-1}\frac{V_\ell}{\sigma^2},$$

where $V_\ell = \sum_{t=1}^{n} A_{\ell,t}A_{\ell,t}^\top$ is the outer product of the feature vectors of the pulled arms in task ℓ and $B_\ell = \sum_{t=1}^{n} A_{\ell,t}Y_{\ell,t}(A_{\ell,t})$ is their sum weighted by their rewards (see Lemma 7 of Kveton et al. (2021) for more details). By Proposition 1, we can calculate the task prior for task s as $P_s = \mathcal{N}(\hat{\theta}_s, \hat{\Sigma}_s + \Sigma_0)$. When $K = d$ and $\mathcal{A}$ is the standard Euclidean basis of $\mathbb{R}^d$, the linear bandit reduces to a K-armed bandit.

Assuming that $\max_{a\in\mathcal{A}} \|a\| \leq 1$ by a scaling argument, the following result holds by an application of our reduction in Section 3.1, and we prove it in Appendix C.1. For a matrix $A \in \mathbb{R}^{d\times d}$, let $\lambda_1(A)$ denote its largest eigenvalue.

Corollary 5.2 (Bayesian Meta Simple Regret, Linear Bandits)**.** *For any* $\delta \in (0, 1]$*, the Bayesian meta simple regret of* `B-metaSRM` *in the setting of Section 5.2 with TS* `alg` *is bounded as* $\mathrm{BSR}(m, n) \leq c_1\sqrt{dm/n} + (m + c_2)\mathrm{SR}_\delta(n) + c_3 dm/n$*, where* $c_1 = O(\sqrt{\log(K/\delta)\log m})$*,* $c_2 = O(\log m)$*, and* c_3 *is a constant in* m *and* n*. Also* $\mathrm{SR}_\delta(n)$ *is the per-task simple regret bounded as* $\mathrm{SR}_\delta(n) \leq c_4\sqrt{\frac{d}{n}} + \sqrt{2\delta\lambda_1(\Sigma_0)}$*, where* $c_4 = O\Big(\sqrt{\log(\frac{K}{\delta})\log n}\Big)$*.*

The first term in the regret is $\tilde{O}(\sqrt{dm/n})$ and represents the price of learning θ_*. The second term is the simple regret of m tasks when θ_* is known and is $\tilde{O}(m\sqrt{d/n})$. The last term is the price of the forced exploration and is negligible, $\tilde{O}(m/n)$. Comparing to the analysis in Basu et al. (2021), we prove a similar bound for `B-metaSRM` with BayesUCB base algorithm in Appendix C.3.

In the frequentist setting, we simplify the setting to $P_* = \mathcal{N}(\theta_*, \sigma_0^2 I_d)$. The case of general covariance matrix for the MAB Gaussian is dealt with in Simchowitz et al. (2021). We extend the results of Simchowitz et al. (2021) for meta-learning to linear bandits. Our estimator of θ_*, namely $\hat{\theta}_s$, is such that $\mathrm{TV}\left(P_{\theta_s} \,||\, P_{\hat{\theta}_*}\right)$ is bounded based on all the observations up to task s. We show that for any $\epsilon, \delta \in (0, 1)$, with probability at least $1 - \delta$ over the realizations of the tasks and internal randomization of the meta-learner, $\hat{\theta}_*$ is close to θ_* in TV distance.

The key idea of the analysis is bounding the regret relative to an oracle. We use Theorem 4 to bound the regret of `f-metaSRM` relative to an oracle $\mathtt{alg}(\theta_*)$ which knows the correct prior. Our analysis and estimator also apply to sub-Gaussian distributions, but we stick to linear Gaussian bandits for readability. Without loss of generality, let $a_1, \ldots, a_d$ be a basis for $\mathcal{A}$ such that $\mathrm{Span}(\{a_1, \ldots, a_d\}) = \mathbb{R}^d$. Resembling Section 5.1, we only need to explore the basis. The exploration strategy, `explore` in Algorithm 2, samples the basis $a_1, \ldots, a_d$ in the first $m_0 \leq m$ tasks. Then the least-squares estimate of θ_* is

$$\hat{\theta}_* := V_{m_0}^{-1}\sum_{s=1}^{m_0}\sum_{i=1}^{d} a_i y_{s,i}\ , \tag{10}$$

where $V_{m_0} := m_0\sum_{i=1}^{d} a_i a_i^\top$ is the outer product of the basis. This gives an unbiased estimate of θ_*. Then we can guarantee the performance of `explore` as follows.

Theorem 6 (Linear Bandits Frequentist Estimator)**.** *In the setting of Section 5.2, for any* ϵ *and* $\delta \in (2e^{-d}, 1)$*, if* $n \geq d$ *and* $m_0 \geq \Big(\frac{d\log(2/\delta)\sum_{i=1}^{d}\sigma_i^2}{2\sigma_0\lambda_d^4(\sum_{i=1}^{d} a_i a_i^\top)\epsilon^4}\Big)^{1/3}$*, then* $\mathrm{TV}(P_{\theta_*} \,||\, P_{\hat{\theta}_*}) \leq \epsilon$ *with probability at least* $1 - \delta$*.*

We prove this by bounding the TV distance of the estimate and correct prior using the Pinsker's inequality. Then the KL-divergence of the correct prior and the prior with parameter $\hat{\theta}_*$ boils down to $\|\theta_* - \hat{\theta}_*\|_2$, which is bounded by the Bernstein's inequality (see Appendix E.4 for the proof).

Now it is easy to bound the frequentist meta simple regret of `f-metaSRM` using the sub-Gaussian version of Corollary 4.1 in Appendix E. We prove the following result in Appendix E.4 by decomposing the simple regret into the relative regret of the base algorithm w.r.t. the oracle.

Corollary 5.3 (Frequentist Meta Simple Regret, Linear Bandits)**.** *In Algorithm 2, let* `alg` *be a TS algorithm and use Eq.* (10) *for estimating the prior parameters with* ${m_0}^3 \geq \left(\frac{d\log(2/\sqrt{\delta})\sum_{i=1}^d \sigma_i^2}{2\sigma_0 \lambda_d^4 (\sum_{i=1}^d a_i a_i^\top)\epsilon^4}\right)$*. Then the frequentist meta simple regret of Algorithm 2 is* $\tilde{O}\left(2m^{1/4}n\,\mathrm{diam}(\mathbb{E}_{\theta_*}[\mu]) + m\frac{d^{3/2}\log K}{\sqrt{n}}\right)$ *with probability at least* $1-\delta$*.*

This bound is $\tilde{O}(m^{1/4}n\|\theta_*\|_\infty + md^{3/2}/\sqrt{n})$, where $\|\cdot\|_\infty$ is the infinity norm. The first term is the price of estimating the prior and the second one is the standard frequentist regret of linear TS for m tasks divided by n, $\tilde{O}(md^{3/2}/\sqrt{n})$. Compared to Corollary 5.2, the above regret bound is looser.

6 Related Work

To the best of our knowledge, there is no prior work on meta-learning for SRM. We build on several recent works on meta-learning for cumulative regret minimization (Bastani, Simchi-Levi, and Zhu 2019; Cella, Lazaric, and Pontil 2020; Kveton et al. 2021; Basu et al. 2021; Simchowitz et al. 2021). Broadly speaking, these works either study a Bayesian setting (Kveton et al. 2021; Basu et al. 2021; Hong et al. 2022), where the learning agent has access to a prior distribution over the meta-parameters of the unknown prior P_*; or a frequentist setting (Bastani, Simchi-Levi, and Zhu 2019; Cella, Lazaric, and Pontil 2020; Simchowitz et al. 2021), where the meta-parameters of P_* are estimated using frequentist estimators. We study both the Bayesian and frequentist settings. Our findings are similar to prior works, that the Bayesian methods have provably lower regret but are also less general when insisting on the exact implementation.

Meta-learning is an established field of machine learning (Thrun 1996, 1998; Baxter 1998, 2000; Finn, Xu, and Levine 2018), and also has a long history in multi-armed bandits (Azar, Lazaric, and Brunskill 2013; Gentile, Li, and Zappella 2014; Deshmukh, Dogan, and Scott 2017). Tuning of bandit algorithms is known to reduce regret (Vermorel and Mohri 2005; Maes, Wehenkel, and Ernst 2012; Kuleshov and Precup 2014; Hsu et al. 2019) and can be viewed as meta-learning. However, it lacks theory. Several papers tried to learn a bandit algorithm using policy gradients (Duan et al. 2016; Boutilier et al. 2020; Kveton et al. 2020; Yang and Toni 2020; Min, Moallemi, and Russo 2020). These works focus on offline optimization against a known prior P_* and are in the cumulative regret setting.

Our SRM setting is also related to fixed-budget *best-arm identification (BAI)* (Gabillon, Ghavamzadeh, and Lazaric 2012; Alieva, Cutkosky, and Das 2021; Azizi, Kveton, and Ghavamzadeh 2022). In BAI, the goal is to control the probability of choosing a suboptimal arm. The two objectives are related because the simple regret can be bounded by the probability of choosing a suboptimal arm multiplied by the maximum gap.

While SRM has a long history (Audibert and Bubeck 2010; Kaufmann, Cappé, and Garivier 2016), prior works on Bayesian SRM are limited. Russo (2020) proposed a TS algorithm for BAI. However, its analysis and regret bound are frequentist. The first work on Bayesian SRM is Komiyama et al. (2021). Beyond establishing a lower bound, they proposed a Bayesian algorithm that minimizes the (Bayesian) per-task simple regret in Eq. (1). This algorithm does not use the prior P_* and is conservative. As a side contribution of our work, we establish Bayesian per-task simple regret bounds for posterior-based algorithms in this setting.

7 Experiments

In this section, we empirically compare our algorithms by their *average meta simple regret* over 100 simulation runs. In each run, the prior is sampled i.i.d. from a fixed meta-prior. Then the algorithms run on tasks sampled i.i.d. from the prior. Therefore, the average simple regret is a finite-sample approximation of the Bayesian meta simple regret. Alternatively, we evaluate the algorithms based on their frequentist regret in Appendix F. We also experiment with a real-world dataset in Appendix F.1.

We evaluate three variants of our algorithms with TS as `alg`; (1) `f-metaSRM` (Algorithm 2) as a frequentist Meta TS. We tune m_0 and report the point-wise best performance for each task. (2) `B-metaSRM` (Algorithm 1) as a Bayesian Meta-learning algorithm. (3) `MisB-metaSRM` which is the same as `B-metaSRM` except that the meta-prior mean is perturbed by uniform noise from $[-50, 50]$. This is to show how a major meta-prior misspecification affects our Bayesian algorithm. The actual meta-prior is $\mathcal{N}(0, \Sigma_q)$.

We do experiments with Gaussian rewards, and thus the following are our baseline for both MAB and linear bandit experiments. The first baseline is `OracleTS`, which is TS with the correct prior $\mathcal{N}(\theta_*, \Sigma_0)$. Because of that, it performs the best in hindsight. The second baseline is agnostic `TS`, which ignores the structure of the problem. We implement it with a prior $\mathcal{N}(\mathbf{0}_K, \Sigma_q + \Sigma_0)$, since μ_s can be viewed as a sample from this prior when the task structure is ignored. Note that Σ_q is the meta-prior covariance in Section 5.2.

The next set of baselines are state-of-the-art BAI algorithms. As mentioned in Section 6, the goal of BAI is not SRM but it is closely related. A BAI algorithm is expected to have small simple regret for a single task. Therefore, if our algorithms outperform them, the gain must be due to meta-learning. We include sequential halving (`SH`) and its linear variant (`Lin-SH`), which are special cases of `GSE` (Azizi, Kveton, and Ghavamzadeh 2022), as the state-of-the-art fixed-budget BAI algorithms. We also include `LinGapE` (Xu, Honda, and Sugiyama 2018) as it shows superior SRM performance compared to `Lin-SH`. All experiments have $m = 200$ tasks with $n = 100$ rounds in each. Appendix F

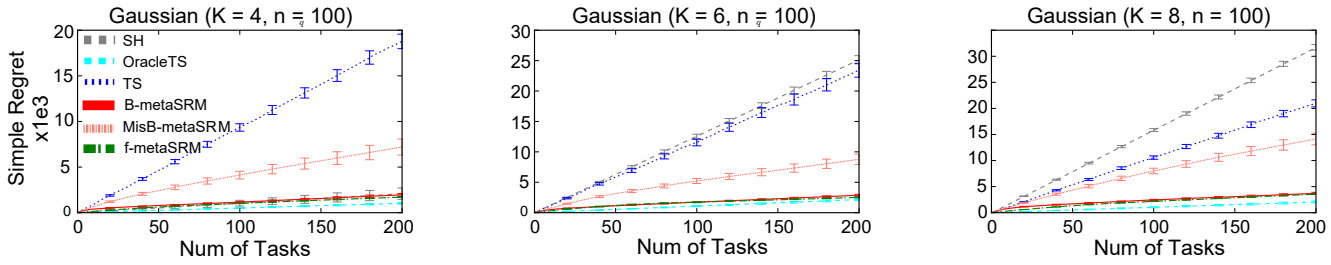

Figure 1: Learning curves for Gaussian MAB experiments. The error bars are standard deviations from 100 runs.

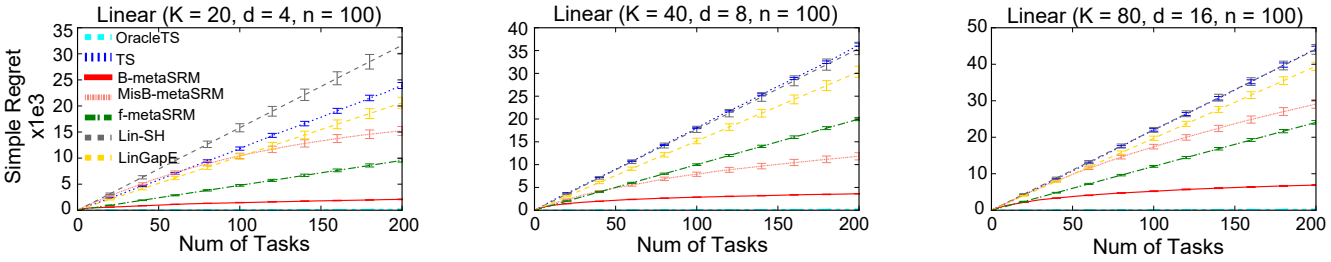

Figure 2: Learning curves for linear Gaussian bandit experiments. The error bars are standard deviations from 100 runs.

describes the experimental setup in more detail and also includes additional results.

7.1 Gaussian MAB

We start our experiments with a Gaussian bandit. Specifically, we assume that $\mathcal{A} = [K]$ are K arms with a Gaussian reward distribution $\nu_s(a; \mu_s) = \mathcal{N}(\mu_s(a), 10^2)$, so $\sigma = 10$. The mean reward is sampled as $\mu_s \sim P_{\theta_*} = \mathcal{N}(\theta_*, 0.1^2 I_K)$, so $\Sigma_0 = 0.1^2 I_K$. The prior parameter is sampled from meta-prior as $\theta_* \sim Q = \mathcal{N}(\mathbf{0}_K, I_K)$, i.e., $\Sigma_q = I_K$.

Fig. 1 shows the results for various values of K. We clearly observe that the meta-learning algorithms adapt to the task prior and outperform `TS`. Both `f-metaSRM` and `B-metaSRM` perform similarly close to `OracleTS`, which confirms the negligible cost of learning the prior as expected in our bounds. We also note that `f-metaSRM` outperforms `MisB-metaSRM`, which highlights the reliance of the Bayesian algorithm on a good meta-prior. `SH` matches the performance of the meta-learning algorithms when $K = 4$. However, as the task becomes harder ($K > 4$), it underperforms our algorithms significantly. For smaller K, the tasks share less information and thus meta-learning does not improve the learning as much.

7.2 Linear Gaussian Bandits

Now take a linear bandit (Section 5.2) in d dimensions with $K = 5d$ arms, the arms are sampled from a unit sphere uniformly. The reward of arm a is distributed as $\mathcal{N}(a^\top \mu_s, 10^2)$, so $\sigma = 10$, and μ_s is sampled from $P_* = \mathcal{N}(\theta_*, 0.1^2 I_d)$, so $\Sigma_0 = 0.1^2 I_d$. The prior parameter, θ_*, is sampled from meta-prior $Q = \mathcal{N}(\mathbf{0}_d, I_d)$, so $\Sigma_q = I_d$.

Fig. 2 shows experiments for various values of d. As expected, larger d increase the regret of all the algorithms. Compared to Section 7.1, the problem of learning the prior is more difficult, and the gap of `B-metaSRM` and `OracleTS` increases. `f-metaSRM` also outperforms `TS`, but it has a much higher regret than `B-metaSRM`. While `MisB-metaSRM` under-performs `f-metaSRM` in the MAB tasks, it performs closer to `B-metaSRM` in this experiment. The BAI algorithms, `Lin-SH` and `LinGapE`, under-perform our meta-learning algorithms and are closer to `TS` than in Fig. 1. The value of knowledge transfer in the linear setting is higher since the linear model parameter is shared by many arms.

Our linear bandit experiment confirms the applicability of our algorithms to structured problems, which shows potential for solving real-world problems. Specifically, the success of `MisB-metaSRM` confirms the robustness of `B-metaSRM` to misspecification.

8 Conclusions and Future Work

We develop a meta-learning framework for SRM, where the agent improves by interacting repeatedly with similar tasks. We propose two algorithms: a Bayesian algorithm that maintains a distribution over task parameters and the frequentist one that estimates the task parameters using frequentist methods. The Bayesian algorithm has superior regret guarantees while the frequentist one can be applied to a larger family of problems.

This work lays foundations for Bayesian SRM and readily extends to reinforcement learning (RL). For instance, we can extend our framework to task structures, such as parallel or arbitrarily ordered (Wan, Ge, and Song 2021; Hong et al. 2022). Our Bayesian algorithm easily extends to tabular and factored MDPs RL (Lu and Van Roy 2019). Also, the frequentist algorithm applies to POMDPs (Simchowitz et al. 2021).

References

Abeille, M.; and Lazaric, A. 2017. Linear Thompson Sampling Revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Agrawal, S.; and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, 39–1. JMLR Workshop and Conference Proceedings.

Agrawal, S.; and Goyal, N. 2013. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, 99–107. PMLR.

Alieva, A.; Cutkosky, A.; and Das, A. 2021. Robust Pure Exploration in Linear Bandits with Limited Budget. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 187–195. PMLR.

Audibert, J.-Y.; and Bubeck, S. 2010. Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, 13 p. Haifa, Israel.

Azar, M. G.; Lazaric, A.; and Brunskill, E. 2013. Sequential Transfer in Multi-Armed Bandit with Finite Set of Models. In *Advances in Neural Information Processing Systems 26*, 2220–2228.

Azizi, M.; Kveton, B.; and Ghavamzadeh, M. 2022. Fixed-budget best-arm identification in structured bandits. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2798–2804.

Bastani, H.; Simchi-Levi, D.; and Zhu, R. 2019. Meta Dynamic Pricing: Transfer Learning Across Experiments. *Management Science*.

Basu, S.; Kveton, B.; Zaheer, M.; and Szepesvári, C. 2021. No Regrets for Learning the Prior in Bandits. *Advances in Neural Information Processing Systems*, 34.

Baxter, J. 1998. Theoretical Models of Learning to Learn. In *Learning to Learn*, 71–94. Springer.

Baxter, J. 2000. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12: 149–198.

Boutilier, C.; Hsu, C.-W.; Kveton, B.; Mladenov, M.; Szepesvari, C.; and Zaheer, M. 2020. Differentiable Meta-Learning of Bandit Policies. In *Advances in Neural Information Processing Systems 33*.

Cella, L.; Lazaric, A.; and Pontil, M. 2020. Meta-Learning with Stochastic Linear Bandits. In *Proceedings of the 37th International Conference on Machine Learning*.

Deshmukh, A. A.; Dogan, U.; and Scott, C. 2017. Multi-Task Learning for Contextual Bandits. In *Advances in Neural Information Processing Systems 30*, 4848–4856.

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Finn, C.; Xu, K.; and Levine, S. 2018. Probabilistic Model-Agnostic Meta-Learning. In *Advances in Neural Information Processing Systems 31*, 9537–9548.

Gabillon, V.; Ghavamzadeh, M.; and Lazaric, A. 2012. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In Pereira, F.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 25, 3212–3220. Curran Associates, Inc.

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

Hoffman, M.; Shahriari, B.; and Freitas, N. 2014. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, 365–374. PMLR.

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

Hsu, C.-W.; Kveton, B.; Meshi, O.; Mladenov, M.; and Szepesvari, C. 2019. Empirical Bayes regret minimization. *arXiv preprint arXiv:1904.02664*.

Kaufmann, E.; Cappé, O.; and Garivier, A. 2016. On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *JMLR*, 17: 1:1–1:42.

Komiyama, J.; Ariu, K.; Kato, M.; and Qin, C. 2021. Optimal Simple Regret in Bayesian Best Arm Identification. *arXiv preprint arXiv:2111.09885*.

Kuleshov, V.; and Precup, D. 2014. Algorithms for Multi-Armed Bandit Problems. *CoRR*, abs/1402.6028.

Kveton, B.; Mladenov, M.; Hsu, C.-W.; Zaheer, M.; Szepesvari, C.; and Boutilier, C. 2020. Differentiable meta-learning in contextual bandits. *arXiv e-prints*, arXiv–2006.

Kveton, B.; wei Hsu, C.; Boutilier, C.; Szepesvari, C.; Zaheer, M.; Mladenov, M.; and Konobeev, M. 2021. Meta-Thompson Sampling. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 5884–5893.

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

Liu, Y.; Devraj, A. M.; Van Roy, B.; and Xu, K. 2022. Gaussian Imagination in Bandit Learning. *arXiv preprint arXiv:2201.01902*.

Lu, X.; and Van Roy, B. 2019. Information-Theoretic Confidence Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 32.

Maes, F.; Wehenkel, L.; and Ernst, D. 2012. Meta-Learning of Exploration/Exploitation Strategies: The Multi-Armed Bandit Case. In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence*, 100–115.

Mason, B.; Jain, L.; Tripathy, A.; and Nowak, R. 2020. Finding All \epsilon-Good Arms in Stochastic Bandits. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 20707–20718. Curran Associates, Inc.

Min, S.; Moallemi, C. C.; and Russo, D. J. 2020. Policy gradient optimization of Thompson sampling policies. *arXiv preprint arXiv:2006.16507*.

Réda, C.; Kaufmann, E.; and Delahaye-Duriez, A. 2021. Top-m identification for linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 1108–1116. PMLR.

Russo, D. 2020. Simple Bayesian Algorithms for Best-Arm Identification. *Operations Research*, 68(6): 1625–1647.

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

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

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

Tripathi, R. C.; Gupta, R. C.; and Gurland, J. 1994. Estimation of parameters in the beta binomial model. *Annals of the Institute of Statistical Mathematics*, 46(2): 317–331.

Vermorel, J.; and Mohri, M. 2005. Multi-Armed Bandit Algorithms and Empirical Evaluation. In *Proceedings of the 16th European Conference on Machine Learning*, 437–448.

Vershynin, R. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wan, R.; Ge, L.; and Song, R. 2021. Metadata-based Multi-Task Bandits with Bayesian Hierarchical Models. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Xu, L.; Honda, J.; and Sugiyama, M. 2018. Fully adaptive algorithm for pure exploration in linear bandits. arXiv:1710.05552.

Yang, K.; and Toni, L. 2020. Differentiable linear bandit algorithm. *arXiv preprint arXiv:2006.03000*.