Bandit Meta-Learning with a Small Set of Optimal Arms

Yasin Abbasi-Yadkori
DeepMind, London, UK
yadkori@deepmind.com

Thang Duong
VinAI Research and VinUniversity, Hanoi, Vietnam
v.thangdn3@vinai.io

Claire Vernade
DeepMind, London, UK
vernade@deepmind.com

András György
DeepMind, London, UK
agyorgy@deepmind.com

Abstract

We study a meta-learning problem where the learner faces a sequence of N multi-armed bandit tasks. Each task is a K-armed bandit problem of horizon T that may be designed by an adversary, but the adversary is constrained to choose the optimal arm of each task in a smaller (but unknown) subset of M arms. An effective strategy is expected to learn the common structure and exploit this knowledge in solving the future tasks. We show an algorithm with a worst-case regret bounded as $\widetilde{O}(N\sqrt{MT}+T\sqrt{KMN})$. We also show that at the cost of an extra $O(\log(N))$ term, the problem can be solved in a computationally efficient way.

1 Introduction

In several real world applications, a learner needs to solve a sequence of tasks. When the tasks share some similarities, the learner may be able to utilize this by sequentially learning the underlying common structure in the tasks and hence perform better on the sequence than solving each task in isolation. In this paper we study a special class of these *meta-learning* problems, where each task is an instance of a multi-armed bandit problem. Such problem settings arise, for example, in recommender systems, where user preferences have some similarities. In such applications, each user represents a different problem/task. One way to model user similarity is to assume that user preferences overlap, and there is a relatively small set of recommendations that can be optimal to any of the users: that is, only a small set of arms can be optimal in any of the bandit problems. In this paper we study such sequences of bandit problems.

Formally, we consider a problem where a learner faces N instances of K-armed bandit tasks sequentially. At the beginning of task $n \in [N]$,[1] the adversary chooses the mean reward function $r_n \in \mathcal{R} = [0,1]^K$. Then, the learner interacts with the K-armed bandit task specified by the reward function r_n for T time steps: whenever the learner plays an action a in the n-th task, it receives a random $[0,1]$-valued reward with expected value $r_n(a)$; we assume that, conditioned on the learner's actions, the rewards are independent (at time step t of task n, the reward received is $r_n(a) + \eta_{n,t}(a)$ where the $\eta_{n,t}(a)$ are independent zero-mean noise variables for all n, t and a). Let $\mathcal{H}_{n,t}$ be the history of the learner's actions and the corresponding rewards up to, but not including, time step t in task n. For this time step, the learner chooses a distribution $\pi_{n,t}$ as a function of $\mathcal{H}_{n,t}$, and samples an action $A_{n,t}$ from $\pi_{n,t}$ to be used in this time step. The learner's policy π is a collection of the mappings $\pi_{n,t}$ (formally, π is a mapping from the set of possible histories to distributions over the

[1] Throughout the paper, for any integer k, we use $[k]$ to denote the set $\{1, 2, \ldots, k\}$, and for any (multi-)set S, $|S|$ denotes the number of distinct elements in S.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

action set $[K]$). Its worst-case expected regret is defined as

$$\text{Regret}_N(\pi, M) = \sup_{r_1,\ldots,r_N \in [0,1]^K} \max_{\substack{a_1,\ldots,a_N \in [K], \\ |\{a_1,\ldots,a_N\}| \le M}} \mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=1}^{T} r_n(a_n) - \sum_{n=1}^{N}\sum_{t=1}^{T} r_n(A_{n,t})\right], \quad (1)$$

where the expectation is over the randomization of the learner. Note that, as usual in bandit problems, we already use the mean reward functions r_n in the regret definition. Importantly, the learner competes with a sequence of arms that has at most M distinct elements. We say a sequence of reward functions $r_1, \ldots, r_N$ is *realizable* if there exists a set of arms of size at most $M \le K$ which contains an optimal arm (achieving reward $\max_a r_n(a)$) for every task n. Otherwise, the problem is called agnostic.

A natural solution to the above bandit meta-learning problem (in the realizable setting) is to use an appropriate base bandit algorithm restricted to the optimal, but unknown, subset of M arms. If the learner believes that the correct subset containing all potentially optimal arms is identified, then a restricted version of the base bandit algorithm can be applied to the next arriving task. Let $B_{T,K}$ be a worst-case regret upper bound for the *base* bandit algorithm (such as UCB, Thompson sampling, or EXP3, see, e.g., [35]) in a K-armed bandit task:

$$\sup_{r_n} \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} r_n(a) - \sum_{t=1}^{T} r_n(A_{n,t})\right] \le B_{T,K}.$$

Note that $B_{T,K} = \widetilde{O}(\sqrt{KT})$[2] for the standard bandit algorithms [35]. When restricted to the correct M-subset (i.e., a subset of size M) of the K arms, the regret of the base algorithm with respect to the best arm in the subset scales as $B_{T,M}$ which is smaller than the basic regret bound $B_{T,K}$ that would be incurred without the restriction. Otherwise, if the chosen subset does not contain the best arm for the next task, the regret on this task can be linear. Therefore, it is important to identify the correct subset quickly.

This way the bandit meta-learning problem can be reduced to a bandit subset-selection problem, where the decision space is the set of M-subsets of $[K]$ and the reward in round n is the maximum of r_n over the chosen subset. As we show in Section 2, an $o(N)$-regret for the bandit subset-selection problem translates to an $O(NB_{T,M}) + o(N)$ bound on Regret_N for the bandit meta-learning problem. Notice that the leading term in the regret is $O(NB_{T,M})$, which would be the regret of a bandit strategy that runs the base algorithm on the correct subset in each task.

This subset-selection problem is a bandit submodular maximization problem, for which Streeter and Golovin [42] presented an algorithm with regret $O(N^{2/3})$. Yet, using a minimax duality argument and the connections between Bayesian and adversarial regret, we show in Section 2.2 that the minimax rate in the bandit subset-selection problem is $O(N^{1/2})$, even in the agnostic setting. Thus the minimax rate in the bandit meta-learning problem is bounded by $\widetilde{O}(N\sqrt{MT} + T\sqrt{KMN})$. This result also shows a regret-minimization procedure in the Bayesian setting (where the bandit tasks are chosen randomly from some unknown distribution and expectation is also taken over this distribution in the definition of the regret), although it does not lead to a computationally efficient solution, and it does not give an actual algorithm in our bandit meta-learning problem.

In turn, in Section 3, we analyze the problem in the realizable setting, and provide two algorithms which provide sample-efficient solutions to the meta-learning problem under some assumptions on the identifiability of the optimal arms in each task, which are satisfied, e.g., if the (mean) rewards of the optimal and suboptimal arms in each task are separated by a large enough gap (as typical in the literature of best arm identification [3]). Our computationally efficient algorithm achieves an $O(NB_{T,M(1+\log N)}) + \widehat{O}(N^{1/2})$ regret, paying a small $\log N$ factor in the regret for computational efficiency, while our other algorithm, whose computational complexity may be exponential in M, achieves the desired $\widetilde{O}(NB_{T,M}) + \widetilde{O}(N^{1/2})$ regret (this is also achieved by our efficient algorithm under the additional assumption that the optimal arm is unique in each task).

Finally, we make a connection between meta-learning problems and partial monitoring games. In fact, it is a common pattern in meta-reinforcement learning benchmarks to assume the existence of certain information gathering actions [45, 20] (also used in our algorithms). Such tasks have a partial

[2]The notation $\widetilde{O}$ hides polylogarithmic terms.

monitoring flavour, although to the best of our knowledge, this connection has not been made explicit yet, and we discuss it in details in Appendix C.

To summarize, we make the following contributions: We introduce the problem of bandit meta-learning with a small set of optimal arms, a simple model that captures many real-world applications. We show a reduction to a bandit subset-selection problem, which is an instance of bandit submodular maximization. By utilizing connections to Bayesian analysis, we show that the minimax regret is of order $O(\sqrt{N})$, which implies $NB_{T,M} + O(\sqrt{N})$ regret in bandit meta-learning. Finally, under a gap condition, we show computationally efficient algorithms that achieve this rate.

2 A reduction to a subset-selection problem

A natural approach to our meta-bandit game is to play in the space of M-subsets: the meta-learner chooses one subset $x_n \in \mathcal{X} = \{x : x \in 2^{[K]}, |x| \leq M\}$ for each task and runs a base bandit algorithm on that subset for T steps and receives an average pseudo-reward[3]

$$\frac{1}{T}\sum_{t=1}^{T} r_n(A_{n,t}) = \max_{a\in x_n} r_n(a) - \frac{1}{T}\left(\sum_{t=1}^{T} \max_{a\in x_n} r_n(a) - r_n(A_{n,t})\right) \doteq f(r_n, x_n) - \varepsilon_n \ ,$$

where $f(r, x) = \max_{a\in x} r(a)$ is the max-reward function for a set of arms x and reward function r and $\varepsilon_n > 0$ is some noise for the meta-learner on its mean reward that corresponds to the average regret incurred by the base algorithm on the subset bandit problem. For any reasonable base algorithm, as discussed in the introduction, we have $\mathbb{E}[\varepsilon_n] \leq B_{T,M}/T = \widehat{O}(\sqrt{K/T})$. To simplify notation, we denote the max-reward function in task n by $f_n = f(r_n, \cdot)$, and the family of all such max-reward functions by $\mathcal{F}$. Now we can easily bound $\text{Regret}_N(\pi, M)$ in the meta-bandit problem as

$$\begin{aligned}
&\text{Regret}_N(\pi, M) \\
&= \sup_{r_1,\ldots,r_N\in[0,1]^K} \max_{\substack{a_1,\ldots,a_N\in[K],\\ |\{a_1,\ldots,a_N\}|\leq M}} \mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=1}^{T}(r_n(a_n) - r_n(A_{n,t}))\right] \\
&= \sup_{r_1,\ldots,r_N\in[0,1]^K} \max_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{n=1}^{N}\sum_{t=1}^{T}(\max_{a\in x} r_n(a) - r_n(A_{n,t}))\right] \\
&= \sup_{r_1,\ldots,r_N\in[0,1]^K} \max_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{n=1}^{N}\left(\sum_{t=1}^{T}(\max_{a\in x} r_n(a) - \max_{a\in x_n} r_n(a)) + \sum_{t=1}^{T}(\max_{a\in x_n} r_n(a) - r_n(A_{n,t}))\right)\right] \\
&= \sup_{f_1,\ldots,f_N\in\mathcal{F}} \max_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{n=1}^{N}(T(f_n(x) - f_n(x_n)) + T\varepsilon_n)\right] \\
&= T\cdot \sup_{f_1,\ldots,f_N\in\mathcal{F}} \max_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{n=1}^{N} f_n(x) - \sum_{n=1}^{N} f_n(x_n)\right] + NB_{T,M} \ , \qquad (2)
\end{aligned}$$

where the last inequality holds by $T\mathbb{E}[\varepsilon_n] \leq B_{T,M}$.

Therefore, the bandit meta-learning problem can be reduced to minimizing a notion of regret where in task n, the learner chooses a subset $x_n \in \mathcal{X}$ and observes $f_n(x_n) - \varepsilon_n$: From this inequality emerges a new notion of minimax *meta-regret* that corresponds to the loss of the meta-learner ρ playing in the action space $\mathcal{X}$ of M-subsets:[4]

$$\text{Regret}_N^{\text{meta}} = \inf_{\rho} \sup_{(f_n)_{n=1}^N} \max_{x\in\mathcal{X}} \mathbb{E}_{x_n\sim\rho}\left[\sum_{n=1}^{N} f_n(x) - \sum_{n=1}^{N} f_n(x_n)\right] \ . \qquad (3)$$

[3] We call this quantity pseudo-reward since it considers the mean reward functions but the actual random actions. The observations of the learner are the actual noisy reward values.

[4] ρ is a sequence of mappings $\rho_{n,t}$ from the history $\bar{\mathcal{H}}_n = ((x_j, \sum_t [r_j(A_{j,t}) + \eta_{j,t}(A_{j,t})])_{j=1..n-1})$ to a meta-action $x_n \in \mathcal{X}$, as opposed to π which directly maps $\mathcal{H}_{n,t}$ to $A_{n,t}$ at each step t of each task n (note that $\bar{\mathcal{H}}_n$ is a function of $\mathcal{H}_{n,T}$).

This reduction to a bandit problem allows us to leverage the literature on bandit optimization to get a bound on $\text{Regret}_N^{\text{meta}}$ that translates into a bound on Regret_N. In the rest of this section, we focus on the bandit subset-selection problem (3). A "task" is called a "round" to emphasize that it is one round of the bandit problem over the subsets (this is similar to earlier formulations of meta-learning in the full-information setting, e.g. [16, 29, 28]).

2.1 Bandit submodular maximization

The problem of subset selection with regret criteria (3) is an instance of online submodular maximization. Streeter and Golovin [42] study the problem in four different settings: the full-information setting where the function f_n is fully observed at the end of round n; the priced feedback model where the learner can observe f_n by paying a price C and the price is added to the total regret; a partially transparent model where values of f_n for some subsets are revealed, and the bandit setting where only $f_n(x_n)$ is observed.

The priced feedback model is similar to problems where the best arms can be identified in every task, which we study in Section 3. While Streeter and Golovin [42] show an algorithm with a regret bounded as $O(N^{2/3})$ for the general priced feedback model, we show an algorithm with $O(N^{1/2})$ regret in the bandit subset-selection problem under the aforementioned identifiability conditions and realizability. Radlinski et al. [38] provide a similar algorithm to that of Streeter and Golovin [42] for a particular ranking problem, with the more informative partially transparent feedback model, and also obtain an $O(N^{1/2})$ regret bound.

2.2 An information-theoretic analysis of bandit subset selection

In this section, we analyze the minimax meta-regret in the subset-selection problem, which in turn provides a bound on the minimax regret in our bandit meta-learning problem. These results apply to the agnostic case, as well, not only the realizable setting. We start with a bound on the meta-regret.

Theorem 2.1. *The minimax meta-regret (3) is bounded as* $\text{Regret}_N^{\text{meta}} \leq \sqrt{KN \log(|\mathcal{X}|)}$.

Using decomposition (2) and Theorem 2.1, since $|\mathcal{X}| \leq K^M$, the minimax regret of the bandit meta-learning problem can be bounded as follows:

Corollary 2.1.1. *The minimax regret of the bandit meta-learning problem with M optimal arms is bounded as* $\text{Regret}_N \leq NB_{T,M} + T\sqrt{KMN \log K}$.

The result shows that the $O(N\sqrt{KT})$ regret of a naive algorithm (using a minimax optimal K-arm bandit algorithm) can be improved to $O(N\sqrt{MT} + T\sqrt{KMN \log K})$. The proof of Theorem 2.1, given in Appendix A, is non-constructive (it does not provide an algorithm achieving the minimax regret). It is based on a recent technique developed for the adversarial convex bandit problem [10, 9, 33], which reduces the analysis of the minimax regret to the *Bayesian* regret, and relies on the main idea of Russo and Van Roy [41] who designed a strategy that balances instantaneous regret and gathering information about the underlying problem. Using results from Lattimore and György [34], we can construct a policy that minimizes the frequentist minimax regret, although the resulting strategy would be computationally very inefficient.

On the other hand, we can construct computationally efficient algorithms for the realizable setting under some mild additional assumptions on the identifiability of the optimal arms (which are satisfied, e.g., under some appropriate gap conditions on the reward functions). This is explored in the next section.

3 An efficient solution under an identifiability condition

In this section, we show an efficient algorithm assuming that the learner has access to an exploration method that reveals optimal actions.

Assumption 1 (Efficient Identification)**.** *There exists a set of M arms that has non-empty intersection with the set of optimal arms in each round. Furthermore, the learner has access to a best-arm-identification (BAI) procedure that for some $\delta, \Delta \in [0, 1]$, with probability at least $1 - \delta/N$, identifies the set of optimal arms if executed in a task (for at most T steps).*

This assumption is satisfied in particular when the optimality gap between the best and second-best arms is large enough given T, N and δ such that there exists a BAI algorithm that returns the best arm after T steps with probability at least $1 - \delta/N$. For example, if for any task n with optimal arms $x_n^* \subset [K]$, we have $r_n(a_n^*) - \max_{a \notin x_n^*} r_n(a) \geq \Delta$ for all $a_n^* \in x_n^*$ (note that $r_n(a_n^*)$ is the same for all a_n^*) for some $\Delta = \Theta(\sqrt{K \log(N/\delta)/T})$, a properly tuned phased elimination (PE)[5] procedure [3] returns the set of optimal arms with probability at least $1 - \delta/N$. The cumulative worst-case regret of PE in a task with K arms has regret $B'_{T,K} = \Theta(B_{T,K})$ [3], see Appendix E for details. With a slight abuse of notation, in this section we use $B_{T,K}$ to denote $\max\{B_{T,K}, B'_{T,K}\}$.

Note that the Efficient Identification assumption requires the BAI procedure to return only optimal arms. This choice is made to keep the analysis simple. The analysis can be extend trivially to allow the returned set to be all arms with sub-optimality gap smaller than $\Theta(\sqrt{M \log(N/\delta)/T})$; we discuss this extension in more details during the anlysis.

We construct a learning algorithm that disentangles exploration (EXR) and exploitation (EXT) at a meta-level: at the beginning of each task n, a meta-decision $E_n \in \{\text{EXR}, \text{EXT}\}$ is taken that conditions the learning during the task. In the exploration mode ($E_n = \text{EXR}$), the learner executes a BAI algorithm on all arms, and (by Assumption 1) observes, with high probability, the set of optimal actions x_n^* chosen by the adversary. The price of this information is a large regret denoted by $\mathbf{C}_{\texttt{info}}$. As discussed above, for a properly tuned PE, we can define $\mathbf{C}_{\texttt{info}} = B_{T,K} > B'_{T,M}$. So, since we aim for $\text{Regret}_N \leq \widetilde{O}(NB_{T,M}) + o(N)$, we should keep the number of EXR calls small. In an exploitation mode ($E_n = \text{EXT}$), the learner executes a base bandit algorithm on a chosen subset x_n, constructed using the previously identified optimal actions $\mathcal{I}_n = \bigcup_{j<n:E_j=\text{EXR}} x_j^*$.

Let s_n be the size of x_n. If $x_n \cap x_n^* \neq \emptyset$, the regret of the base algorithm is bounded by $\mathbf{C}_{\texttt{hit}} = B_{T,s_n}$. Otherwise, since the performance gap between the optimal arms and the arms in x_n can be arbitrary, the regret in the task can be as large as $\mathbf{C}_{\texttt{miss}} = T$. Note that to keep $\mathbf{C}_{\texttt{hit}}$ small, the subset x_n should be as small as possible. Ideally, x_n should be a subset of size M that has non-empty overlap with all members of $\mathcal{I}_n$. However, the problem of finding such x_n is the hitting set problem, which is known to be NP-Complete [21].

Although the exact hitting set problem is computationally hard, a simple greedy algorithm can be used to get an approximate solution efficiently (see, e.g, [42]): it has polynomial computation complexity and achieves an approximation ratio of $1 + \log N$, meaning that it finds a subset of size at most $M(1 + \log N)$ that contains an optimal action for each task. We say an action $a \in [K]$ covers task j if $a \in x_j^*$. The greedy method, denoted by GREEDY, starts with an empty set and at each stage, it adds the action that covers the largest number of uncovered tasks in $\mathcal{I}_n$, until all tasks are covered.

The resulting method, which collects optimal arms in a bottom-up fashion is version `G-BASS` in Algorithm 1. The description also includes a second algorithm, called `E-BASS`, that is based on an elimination procedure. `E-BASS` essentially differs from `G-BASS` in how the information gathered through exploration is used: in `E-BASS`, the learner maintains an active set of possible M-subsets compatible with the observations collected in EXR rounds so far, and all subsets that are inconsistent with prior observations in $\mathcal{I}_n$ are eliminated. In its EXT mode, a subset is selected uniformly at random from the set of active subsets. As we will show, this algorithm improves the regret by a factor of $\log N$, although it is not computationally efficient.

The analysis of `G-BASS` depends on the study of the *cost-to-go* function of the following simple game between a learner and an environment. At each round n, each player has two actions: The learner may choose $E_n \in \{\text{EXR}, \text{EXT}\}$, and its action distribution is characterized by $p_n = P(E_n = \text{EXR})$. The environment may choose a best arm a_n^* that the learner already knows about (i.e, $a_n^* \in x_n$, and x_n is the choice of `G-BASS` for task n) or choose an optimal arm set x_n^* so that $x_n^* \cap x_n = \emptyset$. We denote $q_n = P(x_n^* \cap x_n = \emptyset)$. The cost associated with EXR is always $\mathbf{C}_{\texttt{info}}$, and the learner gets to know the identity of the optimal arms, the cost of EXT is $\mathbf{C}_{\texttt{hit}}$ if the environment chooses an existing optimal arm, and the $\mathbf{C}_{\texttt{miss}}$ if it is a new arm (it is easy to see that one can construct reward functions resulting in the same regret bounds for task n for `G-BASS`, e.g., in the last case the mean rewards of an unobserved arm is 1 and all the other rewards are 0). This implies that the worst-case regret of `G-BASS` can be bounded by the total cost of the learner in this simple game, under the assumption that

[5] Improved UCB in [3] runs exponentially growing phases and maintains an active set of arms with gaps twice smaller in every step.

Algorithm 1: BASS: Bandit Subset Selection for Meta-Learning

Input: Base (efficient K-armed bandit algorithm), best arm identification algorithm BAI, EXR probabilities p_n with $p_1 = 1$, subset size M;
Option: Greedy G-BASS highlighted (G), Elimination-based E-BASS highlighted (E);
Initialize: Let (G) $\mathcal{I}_0 = \emptyset$; (E) $\mathcal{X}_0$ be the set of all M-subsets of $[K]$.;
for $n = 1, 2, \ldots, N$ **do**
  With probability p_n, let $E_n =$ EXR; otherwise let $E_n =$ EXT;
  if $E_n =$ EXR **then**
    Run BAI on all arms and observe the best arms x_n^* of this task;
    (G) Update $\mathcal{I}_n = \mathcal{I}_{n-1} \cup x_n^*$;
    (E) Let $\mathcal{X}_n = \{x \in \mathcal{X}_{n-1} : x \cap x_n^* \neq \emptyset\}$ be all elements of $\mathcal{X}_{n-1}$ with non-empty intersection with x_n^*;
  else
    (G) Find x_n using GREEDY such that $x_n \cap x \neq \emptyset$ for all $x \in \mathcal{I}_n$;
    (E) Sample x_n uniformly at random from $\mathcal{X}_n$;
    Play Base algorithm on x_n ;
  end
end

the best-arm-identification can be performed with probability 1 (i.e. $\delta = 0$ in Efficient Identification assumption). The learner is a (randomized) function of its current knowledge stored in $\mathcal{I}_n$, hence we can easily define and write the minimax cost-to-go function as

$$\begin{aligned} V_N(\mathcal{I}_N) &= 0 \quad \text{and for } n < N, \\ V_n(\mathcal{I}_n) &= \min_p \max_q \{p\mathbf{C}_{\texttt{info}} + q(1-p)\mathbf{C}_{\texttt{miss}} + (1-q)(1-p)\mathbf{C}_{\texttt{hit}} \\ &\qquad + (1-pq)V_{n+1}(\mathcal{I}_n) + pqV_{n+1}(\mathcal{I}_{n+1})\}\,, \end{aligned} \tag{4}$$

where the last equality comes from the fact that when the environment reveals a new action (happens with probability q) and the learner explores (with probability p), its current knowledge set $\mathcal{I}_n$ is incremented. The optimal cost-to-go function V_n above corresponds to the case of $\delta = 0$ in Assumption 1, and $V_0(\emptyset)$ gives the minimax regret for the family of algorithms with the limited choice described. Therefore, when the BAI algorithm is successful almost surely, $\text{Regret}_N \leq V_0(\emptyset)$. For $\delta > 0$, using a union bound, with probability $1 - \delta$, the set of best arms is identified properly in every round, and the maximum regret when this does not happen is NT, hence in this case

$$\text{Regret}_N \leq V_0(\emptyset) + \delta NT.$$

Selecting $\delta = 1/(NT)$ ensures that the last term is negligible compared to the first one. Finally, if the BAI algorithm only returns a set of approximately optimal arms satisfying $r_n(a) \geq r_n(a_n^*) - \Delta$ for all arms a selected, the meta-regret can be bounded trivially as $\text{Regret}_N \leq V_0(\emptyset) + (\delta + \Delta)NT$.

Before deriving $V_0(\emptyset)$ in the general case, we consider the more restricted setting where each task has a unique optimal action.

3.1 Characterizing an optimal policy in the case of unique optimal arms

In this section, we characterize the minimax optimal policy of the learner and the adversary under the condition that there is a unique and identifiable optimal arm in each task.

Assumption 2 (Unique Identification)**.** *Assumption 1 holds, and there is a unique optimal arm in each task.*

Theorem 3.1. *In the simple game,*

$$V_0(\emptyset) \leq NB_{T,M} + M\sqrt{2(\boldsymbol{C}_{info} - \boldsymbol{C}_{hit})(\boldsymbol{C}_{miss} - \boldsymbol{C}_{hit})N}\,.$$

Therefore, under the Unique Identification assumption, the regret of G-BASS can can be bounded, for an appropriate selection of the parameters p_n given in the proof, as

$$\textit{Regret}_N \leq NB_{T,M} + M\sqrt{B_{T,K}TN} + \delta NT\,.$$

Proof. The proof relies on solving the min-max problem in (4). First, we consider the case that the best-arm-identification can be performed with probability 1 (i.e., $\delta = 0$ in the efficient identification assumption). From symmetry, it is easy to see that $V_n(\mathcal{I}_n)$ only depends on the size of $\mathcal{I}_n$, and not the actual arms in $\mathcal{I}_n$. Therefore, to simplify notation and emphasize the dependence on the number of discovered optimal arms, we use below $V_n(|\mathcal{I}_n|) := V_n(\mathcal{I}_n)$. Let $n_M = \operatorname{argmin}_n\{|\mathcal{I}_n| = M\}$ be the first round when all optimal arms have been discovered. Then from any $n > n_M$, the adversary no longer can reveal new arms ($q = 0$), and the learner should no longer explore ($p = 0$), and so

$$\forall n \geq n_M,\ V_n(M) = (N-n)\mathbf{C}_{\texttt{hit}}.$$

Denoting $s = |\mathcal{I}_n|$, the min-max optimization objective in (4) can be written as

$$\begin{aligned} L(q,p) = \mathbf{C}_{\texttt{hit}} + p(\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}}) + V_{n+1}(s) + q(1-p)(\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}}) \\ - p\left[q^1(V_{n+1}(s) - V_{n+1}(s+1)) + \cdots + q^{M-s}(V_{n+1}(s) - V_{n+1}(M))\right];, \end{aligned}$$

where q^i denotes the probability that the environment reveals i optimal arms in the round, and $q = \sum_{i=1}^{M-s} q^i$. Given that $V_{n+1}(s) - V_{n+1}(s+1) < \cdots < V_{n+1}(s) - V_{n+1}(M)$, the maximizing q is such that $q^i = 0$ for $i > 1$ and $q = q^1$. Using this, the saddle point can be obtained by solving $\partial L(q,p)/\partial q = 0$ and $\partial L(q,p)/\partial p = 0$:

$$p = p_n = \frac{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}}}{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}} + V_{n+1}(s) - V_{n+1}(s+1)}, \qquad q_n = \frac{\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}}}{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}} + V_{n+1}(s) - V_{n+1}(s+1)}.$$

Plugging these values in (4), we get

$$V_n(s) = V_{n+1}(s) + \mathbf{C}_{\texttt{hit}} + \frac{(\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}})(\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}})}{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}} + V_{n+1}(s) - V_{n+1}(s+1)}.$$

Given N and M, the policy of the learner and the adversary can be computed by solving the above recursive equation. Given that for any $s < M$, $V_n(s+1) \geq V_{n+1}(s+1) + \mathbf{C}_{\texttt{hit}}$,

$$V_n(s) - V_n(s+1) \leq V_{n+1}(s) - V_{n+1}(s+1) + \frac{(\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}})(\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}})}{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}} + V_{n+1}(s) - V_{n+1}(s+1)}.$$

Let $G_n(s) = V_n(s) - V_n(s+1) \geq 0$ be the cost difference in state s relative to state $s+1$. We have

$$G_n(s) \leq G_{n+1}(s) + \frac{(\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}})(\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}})}{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}} + G_{n+1}(s)}, \tag{5}$$

and indeed by a telescopic argument,

$$\text{Regret}_N - NB_{T,M} = V_0(0) - V_0(M) = \sum_{s=0}^{M-1}(V_0(s) - V_0(s+1)) = \sum_{s=0}^{M-1} G_0(s).$$

The proof is completed by bounding $G_0(s)$ by backward induction on $n \leq N$:

$$G_{N-n}(s) \leq \sqrt{2(\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}})(\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}})n}.$$

The proof of this inequality relies on standard algebraic manipulations that can be found in Appendix B.1. When the BAI routine returns the best arm with probability at least $1 - \delta/N$, with a simple union bound argument, the probability that $\mathcal{I}_n$ ever contains wrong elements is bounded by δ and the above derivations again hold.

□

Interestingly, the exploration probability p_n increases as $\Theta(1/\sqrt{N-n})$. This might seem counter-intuitive at first as typically the exploration rate decreases in most online learning algorithms. The intuition is that as n gets closer to N, if $s < M$, the adversary does not have much time left to use the remaining budget to make the learner suffer a big cost. Therefore, the adversary needs to increase its probability of choosing a new optimal arm.

3.2 The more general case

In this section, we consider the more general case with potentially multiple optimal arms in each task.

Theorem 3.2. *Let $M' = M(1 + \log N)$. Under the Efficient Identification assumption, the regret of the* G-BASS *algorithm with constant exploration probability $p_n = \sqrt{\frac{MT}{NB_{T,K}}}$ for all n is bounded as*

$$\textit{Regret}_N(\texttt{G-BASS}(p)) \leq NB_{T,M'} + MB_{T,K} + \sqrt{MTB_{T,K}N} + \delta NT \ .$$

The proof is similar in spirit to that of Theorem 3.1 above and is deferred to Appendix B.2.

The next theorem shows that the regret of E-BASS is bounded as $NB_{T,M} + o(N)$, which is smaller than the regret of the G-BASS algorithm by a factor of $\log N$; note, however, that E-BASS is not computationally efficient. The proof of the theorem is in Appendix C.

Theorem 3.3. *Under the Efficient Identification Assumption 1, the regret of the* E-BASS *algorithm is bounded as $\textit{Regret}_N \leq NB_{T,M} + O(T^{3/4}K^{1/4}\sqrt{NM \log K})$.*

Remark 1. *The price for the computational efficiency of the* G-BASS *algorithm is the slightly sub-optimal regret bound of $NB_{T,M(1+\log N)} + o(N)$.*

Remark 2 (Connections with partial monitoring games)**.** *The setting of this section can be viewed more generally as a partial monitoring game. Partial monitoring is a general framework in online learning that disentangles rewards and observations (information). In our bandit meta-learning problem, different actions of the meta-learner (*EXR *and* EXT*) provide different levels of information and have different costs, and the problem can be reduced to a partial monitoring game on $\mathcal{X}$, the set of M-subsets of $[K]$. More details are in Appendix C.*

4 Related Work

Our bandit meta-learning problem is based on subset selection, which connects it to many branches of the online learning and bandit literature.

Slate bandits. The reduction (3) in Section 2.2 is an instance of slate bandit problems with a non-separable cost function [19, 39, 26]. An important subproblem in this field is that of learning a diverse ranking [38], which is indeed a special case of submodular bandit optimization [42]. A key difference though is that in these settings the reward is not fully observed, which does not allow them to use GREEDY directly. Nonetheless, the OG algorithm of Streeter and Golovin [42] is relevant in our setting so we can compare to it in our experiments below.

Meta-Learning and Bandit Meta-Learning. Meta, Multi-Task and Transfer Learning [6, 12, 43] are related machine learning problems concerned with learning a lower dimensional subspace across tasks. In that sense, our work is connected to other theoretical studies [22, 17, 16, 18, 30, 29, 44] though indeed we focus on the bandit learning setting. Various other ways of modelling structure have been proposed and studied in bandit meta-learning. A special case of our problem was studied by Azar et al. [5] where K-armed bandit problems are sampled from a prior over a finite set of tasks. Park et al. [37] consider a continual learning setting where the bandit environment changes under a Lipschitz condition. Kveton et al. [31] observe that the hyperparameters of bandit algorithms can be learned by gradient descent across tasks. Learning regularization for bandit algorithms [32, 13] are also proposed, building on the biased regularization ideas from Baxter [6]. Interestingly, these contextual problems are also connected with latent and clustering of bandit models [36, 23–25].

Non-stationary bandits and very large action spaces. Finally, we highlight that our problem is fundamentally (an easy version of) a structured non-stationary bandit problem [47, 25, 40, 4] where change-points are known. As K grows very large, it is also akin to infinitely many armed bandits [7, 46, 8, 11, 15] and countable-armed bandits [27] though these settings do not have a meta-learning aspect.

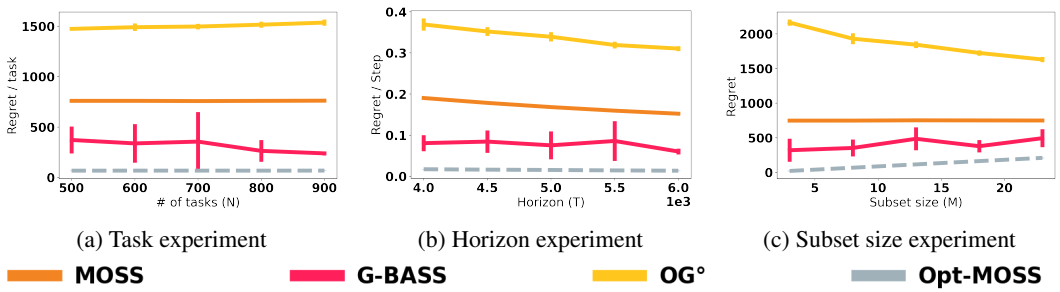
5 Experiments

We empirically validate G-BASS on large scale[6] simulations. The Base algorithm is MOSS [2], and it's called "Opt" when it is played only on the best M-subset, which constitutes an empirical lower

[6] E-BASS is computationally too expensive so we only run it on smaller settings in Appendix D

bound (oracle) to the achievable regret. We also run an agnostic MOSS on all K arms to highlight the improvement from learning the structure. We compare to OG^o[7] ([42], discussed in Section 2.1).

In Figure 1, we demonstrate the impact of four variables on the regret: the number of tasks N (1a), the horizon in each task T (1b) and the optimal susbset size M (1c). To do so, we fix a default setting $(N, T, K, M) = (500, 4000, 80, 8)$ and for each experiment we let one of these parameters vary and observe the regret. In all settings, G-BASS outperforms all methods with a regret close to that of the oracle Opt-MOSS (1c). As opposed to G-BASS, OG^o does not start meta-learning in the limited $N = 900$ rounds (1a), but it is learning within each round as its regret per step goes down (1b). More experiments are shown and commented in Appendix D.



(a) Task experiment (b) Horizon experiment (c) Subset size experiment

Figure 1: Compared regrets as a function of N (1a), T (1b), M (1c), under Assumption 2. Default setting: 500 tasks, 80 arms, horizon = 4000, and subset size = 8. Opt-MOSS plays MOSS on the optimal subset (oracle). G-BASS is near-optimal on all tasks, OG^o has not started to meta-learn yet. Error bars are ± 1 standard deviation, computed over 5 independent runs. We run 4 experiments in parallel on 8 CPUs (i7-9700K CPU @ 3.60GHz) and the total runtime for the three sets of experiments is 1h45.

6 Conclusions

We provided an analysis of the online subset selection problem. Our algorithm BASS has two options: one allowing for computational efficiency at the cost of slightly increased regret, and an optimal but computationally inefficient approach. This problem provides a sound simple base to understand other meta-bandit and meta-reinforcement learning settings, where learning the structure has a cost that needs to be taken into account at a meta-level.

We did not discuss the difficult question of adaptivity: M is assumed to be known, and relaxing this assumption is a generally hard statistical problem that is left for further investigations.

References

[1] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari. Improved algorithms for linear stochastic bandits. In *NIPS*, 2011.

[2] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

[3] P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, pages 55—-65, 2010.

[4] P. Auer, Y. Chen, P. Gajane, C.-W. Lee, H. Luo, R. Ortner, and C.-Y. Wei. Achieving optimal dynamic regret for non-stationary bandits without prior information. In *COLT*, 2019.

[5] M. G. Azar, A. Lazaric, and E. Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In *NIPS*, 2013.

[6] J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

[7] The subscript ^{o} stands for opaque. The algorithm takes a hyperparameter $\gamma \in (0, 1)$ and we optimized it for each experiment to improve the performance.

[7] D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, and L. A. Shepp. Bandit problems with infinitely many arms. *The Annals of Statistics*, 1997.

[8] T. Bonald and A. Proutiere. Two-target algorithms for infinite-armed bandits with Bernoulli rewards. In *In Advances in Neural Information Processing Systems*, 2013.

[9] S. Bubeck and R. Eldan. Exploratory distributions for convex functions. *Mathematical Statistics and Learning*, 2018.

[10] S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *COLT*, 2015.

[11] A. Carpentier and M. Valko. Simple regret for infinitely many armed bandits. In *ICML*, 2015.

[12] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[13] L. Cella, A. Lazaric, and M. Pontil. Meta-learning with stochastic linear bandits. *Arxiv*, 2020.

[14] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[15] H. P. Chan and S. Hu. Infinite arms bandit: Optimality via confidence bounds, 2020.

[16] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil. Learning to learn around a common mean. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[17] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil. Incremental learning-to-learn with statistical guarantees, 2018.

[18] G. Denevi, C. Ciliberto, R. Grazzi, and M. Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[19] M. Dimakopoulou, N. Vlassis, and T. Jebara. Marginal posterior sampling for slate bandits. In *IJCAI*, 2019.

[20] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[21] U. Feige, L. Lovász, and P. Tetali. Approximating min sum set cover. *Algorithmica*, 40(4): 219–234, 2004.

[22] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning, 2018.

[23] C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[24] J. Hong, B. Kveton, M. Zaheer, Y. Chow, A. Ahmed, and C. Boutilier. Latent bandits revisited. In *NeurIPS*, 2020.

[25] J. Hong, B. Kveton, M. Zaheer, Y. Chow, A. Ahmed, M. Ghavamzadeh, and C. Boutilier. Non-stationary latent bandits. *arXiv*, 2020.

[26] S. Kale, L. Reyzin, and R. E. Schapire. Non-stochastic bandit slate problems. In *NIPS*, 2010.

[27] A. Kalvit and A. Zeevi. From finite to countable-armed bandits. In *Conference on Neural Information Processing Systems*, 2020.

[28] M. Khodak, M.-F. Balcan, and A. Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[29] M. Khodak, M.-F. Balcan, and A. Talwalkar. Provable guarantees for gradient-based meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 424–433, 2019.

[30] W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh. Meta-learning for mixed linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[31] B. Kveton, M. Mladenov, C.-W. Hsu, M. Zaheer, C. Szepesvári, and C. Boutilier. Differentiable meta-learning in contextual bandits. *arXiv:2006.05094v1*, 2020.

[32] B. Kveton, M. Konobeev, M. Zaheer, C. wei Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-thompson sampling. *Arxiv*, 2021.

[33] T. Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. 2020.

[34] T. Lattimore and A. György. Mirror descent and the information ratio, 2020.

[35] T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2020.

[36] O.-A. Maillard and S. Mannor. Latent bandits. In *ICML*, 2014.

[37] H. Park, S. Shin, K.-S. Jun, and J. Ok. Transfer learning in bandits with latent continuity. *arXiv*, 2021.

[38] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML*, 2008.

[39] J. Rhuggenaath, A. Akcay, Y. Zhang, and U. Kayma. Algorithms for slate bandits with non-separable reward functions. *arXiv*, 2020.

[40] Y. Russac, C. Vernade, and O. Cappé. Weighted linear bandits for non-stationary environments. In *NeurIPS*, 2019.

[41] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *NIPS*, 2014.

[42] M. Streeter and D. Golovin. An online algorithm for maximizing submodular functions. In *Tech Report. CMU.*, 2007.

[43] S. Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646. MORGAN KAUFMANN PUBLISHERS, 1996.

[44] N. Tripuraneni, C. Jin, and M. I. Jordan. Provable meta-learning of linear representations. *Arxiv*, 2021.

[45] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[46] Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. In *NIPS*, 2008.

[47] C.-Y. Wei and H. Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. *arXiv*, 2021.

[48] J. Yang, W. Hu, J. D. Lee, and S. S. Du. Provable benefits of representation learning in linear bandits. *arXiv*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] **This is a theory paper and we expect no specific societal consequences.**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] **The code for simulations is attached in the supplemental**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] **Everything is reported in the caption of Figure 1**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] **Close to NA but we indicate runtimes in the caption of Figure 1. 8 CPUs i7-9700K CPU @ 3.60GHz**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Theorem 2.1

The proof of the theorem follows the techniques introduced for convex bandits [10, 33]. Since $\mathcal{X}$ in our case is not convex, we cannot directly recycle those results. However, the special structure of our problem makes the derivations simpler, although some tweaks are necessary.

Theorem A.1. *Let $\alpha, \beta \in \mathbb{R}$ be non-negative and for any $f \in \mathcal{F}$, let $f_* = \max_x f(x)$. Suppose that for any g in the convex hull of $\mathcal{F}$ and distribution μ on $\mathcal{F}$, there exists a probability distribution π on $\mathcal{X}$ such that*

$$\int_{\mathcal{F}} f_* d\mu(f) - \sum_{x\in\mathcal{X}} \pi(x) g(x) \le \alpha + \sqrt{\beta \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \pi(x) d\mu(f)} \,. \tag{6}$$

Then the minimax meta-regret is bounded as

$$\mathit{Regret}_N^{\mathit{meta}} \le \alpha N + \sqrt{\beta N \log(|\mathcal{X}|)} \,.$$

Proof. Let $x^* = \operatorname{argmax}_{x\in\mathcal{X}} \sum_{n=1}^N f_n(x)$ be the global best M-subset and $\Delta(\mathcal{F})$ be the space of probability distributions over the function space $\mathcal{F}$. By the minimax theorem,

$$\mathrm{Regret}_N^{\mathrm{meta}} = \inf_\rho \sup_{(f_n)_{n=1}^N} \mathbb{E}_\rho \left(\sum_{n=1}^N f_n(x^*) - \sum_{n=1}^N f_n(x_n) \right) = \sup_{\nu\in\Delta(\mathcal{F}^N)} \inf_\rho \mathbb{E}_{\nu,\rho} \left(\sum_{n=1}^N f_n(x^*) - \sum_{n=1}^N f_n(x_n) \right) \,.$$

On the left-hand side, the expectation is over the randomization in the learning algorithm ρ over the sequence $x_1, \dots, x_N$, and on the right-hand side the expectation is over distribution ν over the sequence $f_1, \dots, f_N$ and the randomization in the learning strategy ρ. Given that the expression on the right-hand side is an expectation with respect to a distribution ν, it is called Bayesian regret and is denoted by $\mathrm{BayesReg}_N = \mathbb{E}_\nu \left(\sum_{n=1}^N f_n(x^*) - \sum_{n=1}^N f_n(x_n) \right)$, and the distribution ν is called the prior.

For any $x, y \in \mathcal{X}$ and n, let

$$\begin{aligned} f_{n,y}(x) &= \mathbb{E}\left[f_n(x) | x^* = y, \nu, f_1(x_1), \dots, f_{n-1}(x_{n-1}) \right] \,, \\ g_n(x) &= \mathbb{E}\left[f_n(x) | \nu, f_1(x_1), \dots, f_{n-1}(x_{n-1}) \right] \,. \end{aligned}$$

$f_{n,y}$ is not in $\mathcal{F}$, so the following argument is not valid. Now define the distributions μ_n such that $\mu_n(\{f_{n,y}\}) = \mathbb{E}[\mathbf{1}(x^* = y) | \nu, f_1(x_1), \dots, f_{n-1}(x_{n-1})]$. By the condition of the theorem, there exists a probability measure π_n on $\mathcal{X}$ such that

$$\int_{\mathcal{F}} f_* d\mu_n(f) - \sum_{x\in\mathcal{X}} \pi_n(x) g_n(x) \le \alpha + \sqrt{\beta \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g_n(x) - f(x))^2 \pi_n(x) d\mu_n(f)} \,.$$

Then, the Bayesian regret of the policy $\{\pi_n\}$ can be bounded as

$$\begin{aligned} \mathrm{BayesReg}_N &= \mathbb{E}_\nu \left[\sum_{n=1}^N \int_{\mathcal{F}} f(x^*) d\mu_n(f) - \sum_{x\in\mathcal{X}} \pi_n(x) g_n(x) \right] \\ &\le \mathbb{E}_\nu \left[\sum_{n=1}^N \int_{\mathcal{F}} f_* d\mu_n(f) - \sum_{x\in\mathcal{X}} \pi_n(x) g_n(x) \right] \\ &\le \alpha N + \mathbb{E}_\nu \left[\sum_{n=1}^N \sqrt{\beta \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g_n(x) - f(x))^2 \pi_n(x) d\mu_n(f)} \right] \,. \end{aligned}$$

Let $v_n = \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g_n(x) - f(x))^2 \pi_n(x) d\mu_n(f)$ be the conditional variance of the right-hand side. By Lemma 5 of Bubeck et al. [10],

$$\mathbb{E}_\nu \left[\sum_{n=1}^N \sqrt{v_n} \right] \le \sqrt{N \mathbb{E}_\nu \left[\sum_{n=1}^N v_n \right]} \le \sqrt{\frac{N}{2} \log(|\mathcal{X}|)} \,.$$

This implies that $\text{BayesReg}_N \le \alpha N + \sqrt{\beta N \log(|\mathcal{X}|)}$ for policy $\{\pi_n\}$ for any prior ν. Therefore the minimax regret can also be bounded as

$$\text{Regret}_N^{\text{meta}} = \sup_\nu \inf_\rho \mathbb{E}[\text{BayesReg}_N] \le \alpha N + \sqrt{\beta N \log(|\mathcal{X}|)} \,.$$

□

To obtain the final bound on the minimax regret, it is left to show that the condition of Theorem A.1 holds. First, we prove the following lemma, which is a slight modification of a similar result of Lattimore [33].

Lemma A.2. *Let $\gamma \in \mathbb{R}$, $\mathcal{F} = \cup_{i=1}^B \mathcal{F}_i$, and g be in the convex hull of $\mathcal{F}$. Assume there exist probability measures $(\pi_i)_{i=1}^B$ on $\mathcal{X}$ such that for all $i \in [B]$, for all $f \in \mathcal{F}_i$,*

$$f_* - \sum_{x\in\mathcal{X}} \pi_i(x) g(x) \le \alpha + \sqrt{\gamma \sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \pi_i(x)} \,.$$

Then, for any distribution μ on $\mathcal{F}$, there exists a probability measure π on $\mathcal{X}$ such that

$$\int_{\mathcal{F}} f_* d\mu(f) - \sum_{x\in\mathcal{X}} \pi(x) g(x) \le \alpha + \sqrt{\gamma B \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \pi(x) d\mu(f)} \,.$$

Proof. Assume without loss of generality that sets $\mathcal{F}_i$ are disjoint and $\mu(\mathcal{F}_i) > 0$. Define probability measure μ_i by $\mu_i(A) = \mu(A \cap \mathcal{F}_i)/\mu(\mathcal{F}_i)$. Let $q_i = \mu(\mathcal{F}_i)$ and $\pi = \sum_{i=1}^B q_i \pi_i$. Then,

$$\begin{aligned}
\int_{\mathcal{F}} f_* d\mu(f) - \sum_{x\in\mathcal{X}} \pi(x) g(x) &= \sum_{i=1}^B q_i \left(\int_{\mathcal{F}} f_* d\mu_i(f) - \sum_{x\in\mathcal{X}} \pi_i(x) g(x) \right) \\
&\le \alpha + \sum_{i=1}^B q_i \sqrt{\gamma \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \pi_i(x) d\mu_i(f)} \\
&\le \alpha + \sqrt{\gamma B \sum_{i=1}^B q_i^2 \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \pi_i(x) d\mu_i(f)} \\
&\le \alpha + \sqrt{\gamma B \sum_{i,j=1}^B q_i q_j \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \pi_i(x) d\mu_j(f)} \\
&= \alpha + \sqrt{\gamma B \int_{\mathcal{F}} \sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \rho(x) d\mu(f)} \,.
\end{aligned}$$

□

We are ready to prove Theorem 2.1.

Proof of Theorem 2.1. We show that the condition of Lemma A.2 is satisfied. Let $\mathcal{F}_i$ be the space of functions parameterized by some reward vector r satisfying $i = \text{argmax}_{j\in[K]} r(j)$. We clearly have $B = K$ such disjoint function spaces. Notice that for any $f \in \mathcal{F}_i$ specified by the reward function r_f,

$$r_f(i) = \max_{j\in[K]} r_f(j) = \max_x \max_{j\in x} r_f(j) = \max_x f(r_f, x) = f_* \,.$$

Let each probability measure π_i be the uniform distribution on all $x \in \mathcal{X}$ such that $i \in x$. Therefore, for any x in the support of π_i and any $f \in \mathcal{F}_i$,

$$f(x) = \max_{j\in x} r_f(j) = r_f(i) = f_* \,.$$

Therefore

$$f_* - \sum_{x\in\mathcal{X}} \pi_i(x) g(x) = \sum_{x\in\mathcal{X}} \pi_i(x)(f(x) - g(x)) \\ \leq \sum_{x\in\mathcal{X}} \pi_i(x)|f(x) - g(x)| \\ \leq \sqrt{\sum_{x\in\mathcal{X}} (g(x) - f(x))^2 \pi_i(x)}\,.$$

Therefore, the condition of Lemma A.2 is satisfied with $\gamma = 1$, and so the condition of Theorem A.1 holds with $\beta = \gamma B = K$ and $\alpha = 0$. □

B Proof details

B.1 Complement to the proof of Theorem 3.1

We are left to prove that

$$G_n(s) \leq G_{n+1}(s) + \frac{(\mathbf{C}_{\mathtt{info}} - \mathbf{C}_{\mathtt{hit}})(\mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}})}{\mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}} + G_{n+1}(s)}, \tag{7}$$

given in (5) implies that for any $n \leq N$,

$$G_n(s) \leq \sqrt{2(\mathbf{C}_{\mathtt{info}} - \mathbf{C}_{\mathtt{hit}})(\mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}})(N-n)}\,. \tag{8}$$

Proof. We proceed by (backward) induction. First, by definition, $G_N(s) = V_N(s) - V_N(s+1) = 0$ for all s, thus (8) holds for $n = N$. Next, assume that (8) holds for $\{N, N-1, \ldots, n+1\}$, and we show that it also holds for n.

Consider positive constants $b \geq a$ and consider the function $h(z) = z + \frac{ab}{b+z}$ defined on $[0, c]$ for some $c > 0$. Then $h'(z) = 1 - ab/(b+z)^2 \geq 0$. Therefore, h is maximized at $z = c$. Since the right-hand side of (7) is of the form $h(G_{n+1}(s))$ with $a = \mathbf{C}_{\mathtt{info}} - \mathbf{C}_{\mathtt{hit}}$ and $b = \mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}}$, which indeed satisfy $b \geq a$. By this argument, the induction assumption, and $0 \leq G_{n+1}(s) \leq \sqrt{ab(N-n-1)}$ by the induction hypothesis, we obtain that

$$\begin{aligned} G_n(s) &\leq \sqrt{2(\mathbf{C}_{\mathtt{info}} - \mathbf{C}_{\mathtt{hit}})(\mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}})(N-n-1)} \\ &\quad + \frac{(\mathbf{C}_{\mathtt{info}} - \mathbf{C}_{\mathtt{hit}})(\mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}})}{\mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}} + \sqrt{2(\mathbf{C}_{\mathtt{info}} - \mathbf{C}_{\mathtt{hit}})(\mathbf{C}_{\mathtt{miss}} - \mathbf{C}_{\mathtt{hit}})(N-n-1)}} \\ &= \sqrt{2ab(N-n-1)} + \frac{ab}{b + \sqrt{2ab(N-n-1)}} \end{aligned} \tag{9}$$

It remains to show that the right-hand side above is bounded from above by $\sqrt{2ab(N-n)}$. This follows since

$$\sqrt{2ab(N-n)} - \sqrt{2ab(N-n-1)} = \frac{\sqrt{2ab}}{\sqrt{N-n} + \sqrt{N-n-1}} \\ = \frac{ab}{\sqrt{ab(N-n)/2} + \sqrt{ab(N-n-1)/2}} \geq \frac{ab}{b + \sqrt{2ab(N-n-1)}}$$

where the last inequality holds because

$$b + \sqrt{2ab(N-n-1)} \geq \left[\sqrt{ab} + \sqrt{ab(N-n-1)/2}\right] + \sqrt{ab(N-n-1)/2} \\ \geq \sqrt{ab(N-n)/2} + \sqrt{ab(N-n-1)/2}$$

(where we used that $1 + \sqrt{z} \geq \sqrt{z+1}$ for $z \geq 0$). Thus, $G_n(s) \leq \sqrt{2ab(N-n}$, proving the induction hypothesis (8) for n.

□

B.2 Proof of Theorem 3.2

Proof. The proof relies on the analysis of the optimization problem defined as in Eq. (4) with a fixed p (no minimzation over p, the exploration probability of the learner at each step). As in the proof of Theorem 3.1, we assume that the best arm identification is successful, and the extension to $\delta \neq 0$ can be done the same way. After some algebraic manipulation, similarly to the proof of Theorem 3.1, the optimization objective can be written as

$$\begin{aligned} L(q) = V_{n+1}(x_n) + B_{T,|x_n|} + p(B_{T,K} - B_{T,|x_n|}) \\ + q\left\{(1-p)(T - B_{T,|x_n|}) - p(V_{n+1}(x_n) - V_{n+1}(x'_n))\right\} \end{aligned}$$

where x'_n is the new greedy subset selected by the learner in time step $n+1$ if $x^*_n \cap x_n = \emptyset$ and the learner chooses to explore at time n (we use the notation x'_n instead of x_{n+1} to emphasize that this corresponds to the aforementioned choices of the learner and the adversary). Given that L is linear in q, the optimal adversary choice is either $q = 0$ or $q = 1$ (similarly as in Theorem 3.1, it is suboptimal for the adversary to reveal multiple optimal arms). We have

$$q = \begin{cases} 0 & \text{if } (1-p)(T - B_{T,|x_n|}) - p(V_{n+1}(x_n) - V_{n+1}(x'_n)) \leq 0\,, \\ 1 & \text{otherwise} \end{cases}$$

When $q = 0$, $V_n(x_n) = V_{n+1}(x_n) + B_{T,|x_n|} + p(B_{T,K} - B_{T,|x_n|})$, and given that $|x_n| \leq M'$, the total contribution of these rounds to the regret is bounded by

$$NB_{T,M'} + pNB_{T,K}\ .$$

Next consider the rounds where $q = 1$. Among these rounds, consider rounds where the adversary chooses a particular arm $a \in x^*$ and the learner chooses to explore (EXR). This arm is not added to the future EXT subset of the learner if instead another arm is used to cover this round. This means that after at most K such rounds, the learner adds a to the EXT subset. Since the learner's regret in the exploration rounds is $B_{T,K}$, in these rounds the cumulative regret is bounded by $MKB_{T,K}$. Since the random choices made by the learner and the adversary are independent in the same round, the adversary reveals all positions after MK/p such tasks in expectation where the adversary's choice is $q = 1$. If in these rounds the learner chooses to exploit, it can suffer a regret T, leading to a total expected regret of at most MKT/p. Thus, the total regret of rounds with $q = 1$ is bounded by

$$\frac{MK}{p}T + MKB_{T,K}\ .$$

By the choice of $p = \sqrt{\frac{MKT}{NB_{T,K}}}$, we obtain that

$$\text{Regret}_N = V_0(\emptyset) \leq NB_{T,M'} + MKB_{T,K} + 2\sqrt{MKTB_{T,K}N}\ .$$

□

C Partial monitoring and Bandit Meta-Learning

Partial monitoring is a general framework in online learning that disentangles rewards and observations (information). It is a game where the learner has Z actions and the adversary has D actions, and it is characterized by two $Z \times D$ matrices (not observed): matrix C maps the learner's action to its cost given the adversary's choice, and matrix X maps the learner's action to its observation given the adversary's choice. In all generality, we consider bandit meta-learning problems with $Z + 1$ learner actions: an EXR action that provides information for a cost $\mathbf{C}_{\texttt{info}}$, and Z other actions that do not provide information but have a hidden cost $\mathbf{C}_{\texttt{hit}}$ or $\mathbf{C}_{\texttt{miss}}$ depending on whether the chosen action had low or high cost respectively.

As defined in the introduction, a bandit subset selection problem is realizable when there is a subset of size M that contains an optimal arm in all rounds. Otherwise, the problem is called agnostic.

In our bandit subset selection problem, $Z = \binom{K}{M} \leq K^M$ and the adversary can have up to 2^K choices depending on the realizable or agnostic nature of the problem. We have $D = M$ if the problem is realizable and if the adversary is constrained to picking a unique optimal arm in each

round. For example, let $M = 2$ and $K = 4$. There are $Z + 1 = \binom{4}{2} + 1 = 7$ learner actions and only $D = 2$ possible choices for the adversary

$$\begin{pmatrix} \text{EXR} \\ \{1,2\} = x^* \\ \{1,3\} \\ \{1,4\} \\ \{2,3\} \\ \{2,4\} \\ \{3,4\} \end{pmatrix} \to C = \begin{pmatrix} \mathbf{C}_{\texttt{info}} & \mathbf{C}_{\texttt{info}} \\ \mathbf{C}_{\texttt{hit}} & \mathbf{C}_{\texttt{hit}} \\ \mathbf{C}_{\texttt{hit}} & \mathbf{C}_{\texttt{miss}} \\ \mathbf{C}_{\texttt{hit}} & \mathbf{C}_{\texttt{miss}} \\ \mathbf{C}_{\texttt{miss}} & \mathbf{C}_{\texttt{hit}} \\ \mathbf{C}_{\texttt{miss}} & \mathbf{C}_{\texttt{hit}} \\ \mathbf{C}_{\texttt{miss}} & \mathbf{C}_{\texttt{miss}} \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 2 \\ \perp & \perp \\ \vdots & \vdots \\ \perp & \perp \end{pmatrix}.$$

The symbol $\perp$ is used to denote no observations. We use $C_{i,y}$ to denote the cost of action $i \in \{\text{EXR}, x_1, \ldots, x_Z\}$ when adversary chooses $a \in [D]$. Thanks to this reduction, we can leverage the partial monitoring literature to obtain an algorithm and the according bounds for our problem as well. We detail this process below. Note that using the vocabulary of online learning, the learner's actions are referred to as "experts".

Next, we describe an algorithm based on the Exponentially Weighted Average (EWA) forecaster. The learner estimates the cost matrix by importance sampling when action EXR is chosen. When EXT is chosen, the learner samples an expert according to EWA weights that depend on the estimated cost matrix. The pseudo-code of the method is shown in Algorithm 2.

Algorithm 2: The partial monitoring algorithm

Exploration probability $p \in (0,1)$, learning rate $\eta > 0$, base costs $\mathbf{C}_{\texttt{info}}, \mathbf{C}_{\texttt{hit}}, \mathbf{C}_{\texttt{miss}}$;
for $n = 1, 2, \ldots, N$ **do**
  With probability p, let $E_n = \text{EXR}$ and otherwise $E_n = \text{EXT}$;
  if $E_n = \text{EXR}$ **then**
    Observe the best arms x_n^* of this round and for all $i \in \text{EXT}$ experts, observe cost C_{i,x_n^*} and let $\widehat{C}_n(i) = (C_{i,x_n^*} - \mathbf{C}_{\texttt{hit}})/p$;
    Update exponential weights $Q_{n,i} \propto \exp(-\eta \sum_{\tau=1}^{n} \widehat{C}_n(i))$;
    Suffer cost $\mathbf{C}_{\texttt{info}}$;
  else
    Sample $x_n \sim Q_{n-1}$;
    Suffer (but do not observe) cost $\mathbf{C}_{\texttt{hit}}$ if $x_n^* \cap x_n \neq \emptyset$ and suffer cost $\mathbf{C}_{\texttt{miss}}$ otherwise;
  end
end

To analyze the algorithm, we consider the realizable and agnostic cases. In the realizable case, there a subset of size M that contains an optimal arm in all rounds. In this case, the exponential weights distribution reduces to a uniform distribution over the subsets that satisfy this condition.

Theorem C.1. *Consider the partial monitoring algorithm shown in Algorithm 2. In the agnostic case, with the choice of* $\delta = O\left(\left(\frac{\boldsymbol{C}_{miss}^2 \log Z}{\boldsymbol{C}_{info}^2 N}\right)^{1/3}\right)$ *and* $\eta = O\left(\left(\frac{\log^2 Z}{\boldsymbol{C}_{info}\boldsymbol{C}_{miss}^2 N^2}\right)^{1/3}\right)$*, the regret of the algorithm is bounded as* $O((\boldsymbol{C}_{info}\boldsymbol{C}_{miss}^2 N^2 \log Z)^{1/3})$*. In the realizable case, with the choice of* $\delta = \sqrt{\frac{\boldsymbol{C}_{miss} \log Z}{\boldsymbol{C}_{info} N}}$ *and* $\eta = 1$*, the regret of the algorithm is bounded as* $O(\sqrt{\boldsymbol{C}_{info}\boldsymbol{C}_{miss} N \log Z})$*.*

Proof. Let function $f_n : [Z+1] \times [D] \to \mathbb{R}^{Z+1}$ be defined by

$$f_n(k, X_{k,y})_i = \mathbf{1}\{k = \text{EXR}\}(C_{i,y} - \mathbf{C}_{\texttt{hit}}) .$$

Therefore, $\sum_{k=1}^{Z+1} f_n(k, X_{k,y})_i = C_{i,y} - \mathbf{C}_{\texttt{hit}}$. Let $\delta \in (0,1]$. With probability δ, let $E_n = \text{EXR}$ and otherwise $E_n = \text{EXT}$. Let $C_n(i) = C_{i,Y_n}$. Define cost estimator

$$\widehat{C}_{n,i} = \frac{f_n(E_n, X_{E_n,Y_n})_i}{\delta} = \frac{\mathbf{1}\{E_n = \text{EXR}\}(C_n(i) - \mathbf{C}_{\texttt{hit}})}{\delta} .$$

Let Q_n be the weights of the EWA forecaster defined using the above costs. For any i, we have $\mathbb{E}(\widehat{C}_n(i)) = \mathbb{E}(C_n(i) - \mathbf{C}_{\texttt{hit}})$. Let E_n be the learner's decision in round n, that is either EXR or a subset chosen by EWA, in which case it is denoted by x_n. We have

$$\mathbb{E}(C_n(E_n)) = \delta \mathbf{C}_{\texttt{info}} + (1-\delta)\mathbb{E}(C_n(x_n)) .$$

Let x^* be the optimal subset. By the regret bound of EWA [14],

$$\sum_{n=1}^{N} \widehat{C}_n(x_n) - \sum_{n=1}^{N} \widehat{C}_n(x^*) \leq \frac{\log Z}{\eta} + \frac{\eta}{2} \sum_{n=1}^{N} \|\widehat{C}_n\|_\infty^2 \ .$$

Thus,

$$\begin{aligned}\sum_{n=1}^{N} \mathbb{E}(C_n(E_n)) - \sum_{n=1}^{N} \mathbb{E}(C_n(x^*)) &\leq \mathbf{C}_{\texttt{info}} \sum_{n=1}^{N} \delta + \frac{\log Z}{\eta} + \frac{\eta}{2} \sum_{n=1}^{N} \mathbb{E}(\|\widehat{C}_n\|_\infty^2) \\ &\leq \mathbf{C}_{\texttt{info}} \sum_{n=1}^{N} \delta + \frac{\log Z}{\eta} + \frac{\eta \mathbf{C}_{\texttt{miss}}^2}{2} \sum_{n=1}^{N} \frac{1}{\delta} \ .\end{aligned}$$

With the choice of $\delta = O((\mathbf{C}_{\texttt{miss}}/\mathbf{C}_{\texttt{info}})^{2/3}(\log^{1/3} Z)/N^{1/3})$ and $\eta = O((\log^{2/3} Z)/(\mathbf{C}_{\texttt{miss}}^{2/3}\mathbf{C}_{\texttt{info}}^{1/3}N^{2/3}))$, the regret of the partial monitoring game is bounded as $O(\mathbf{C}_{\texttt{miss}}^{2/3}\mathbf{C}_{\texttt{info}}^{1/3}N^{2/3}\log^{1/3} Z)$. The regret scales logarithmically with the number of experts, and is independent of the number of adversary choices.

Next, we show a fast $O(\sqrt{N})$ rate when the optimal expert always has small cost. More specifically, we assume that $C_n(x^*) = \mathbf{C}_{\texttt{hit}}$ for the optimal expert x^*. The fast rate holds independently of the relative values of $\mathbf{C}_{\texttt{hit}}, \mathbf{C}_{\texttt{info}}$ and $\mathbf{C}_{\texttt{miss}}$. The algorithm can also be implemented efficiently.

Let $\widehat{\ell}_n = \delta_n \widehat{C}_n/\mathbf{C}_{\texttt{miss}}$, which is guaranteed to be in $[0,1]$. Notice that $\sum_{n=1}^{N} \widehat{\ell}_n(x^*) = 0$ as $C_n(x^*) = \mathbf{C}_{\texttt{hit}}$ by assumption. In this case, the regret of EWA is known to be logarithmic:

$$\sum_{n=1}^{N} \widehat{\ell}_n(x_n) - \sum_{n=1}^{N} \widehat{\ell}_n(x^*) = O(\log Z) \ .$$

Thus,

$$\sum_{n=1}^{N} \mathbb{E}(C_n(E_n)) - \sum_{n=1}^{N} \mathbb{E}(C_n(x^*)) \leq \mathbf{C}_{\texttt{info}} \sum_{n=1}^{N} \delta + \frac{\mathbf{C}_{\texttt{miss}} \log Z}{\delta} \ .$$

Therefore, with the choice of $\delta = \sqrt{\frac{\mathbf{C}_{\texttt{miss}} \log Z}{\mathbf{C}_{\texttt{info}} N}}$,

$$\sum_{n=1}^{N} \mathbb{E}(C_n(E_n)) - \sum_{n=1}^{N} \mathbb{E}(C_n(x^*)) \leq O(\sqrt{\mathbf{C}_{\texttt{info}}\mathbf{C}_{\texttt{miss}} N \log Z}) \ .$$

The meta-regret scales logarithmically with the number of experts, and is independent of the number of adversary choices. Given that the optimal expert is known to have small loss in all rounds, the learner can eliminate all other experts. Therefore, the EWA strategy reduces to a uniform distribution over the surviving experts, and this strategy can be implemented efficiently.

□

C.1 Proof of Theorem 3.3

E-BASS is constructed as a special case of the EWA algorithm above, where the sampling distribution at each EXT round is simply the uniform distribution over the surviving experts. The proof of Theorem 3.3 is therefore a direct consequence of the more general analysis done for the EWA forecaster in Theorem C.1 above.

Proof. The BAI algorithm might return a number of extra arms in addition to the optimal arm. However, since with high probability the optimal arm is always in the surviving set, the cost estimate for the optimal subset is always zero, and costs of all other subsets are under-estimated. Therefore, if x_n is the expert (subset) selected in round n and x^* is the optimal subset, by fast rates of the previous section,

$$\sum_{n=1}^{N} \mathbb{E}(C_n(x_n)) - \sum_{n=1}^{N} \mathbb{E}(C_n(x^*)) \leq O(\sqrt{\mathbf{C}_{\texttt{info}}\mathbf{C}_{\texttt{miss}} N \log Z}) \ .$$

Given that with high probability the optimal arm is always in the surviving set and therefore $C_n(x^*) = \mathbf{C}_{\texttt{hit}}$,

$$\begin{aligned}\text{Regret}_N &= \sum_{n=1}^{N} \mathbb{E}\left(TR_n(a_n^*) - \sum_{t=1}^{T} R_n(A_{n,t})\right) \le \sum_{n=1}^{N} \mathbb{E}(C_n(x_n)) \le N\mathbf{C}_{\texttt{hit}} + O(\sqrt{\mathbf{C}_{\texttt{info}}\mathbf{C}_{\texttt{miss}} N \log Z}) \\ &= N\sqrt{MT} + O(\sqrt{\mathbf{C}_{\texttt{info}}\mathbf{C}_{\texttt{miss}} N \log Z}) \\ &= N\sqrt{MT} + O(T^{3/4}K^{1/4}\sqrt{NM\log(K)}\,,\end{aligned}$$

where the first inequality holds by the fact that $\mathbb{E}(C_n(x_n))$ is an upper bound on the regret for task n. □

C.2 Other meta bandit partial monitoring problems

A number of other bandit meta-learning problems can also be viewed as partial monitoring games. As an example, consider the following meta-learning problem where all bandit tasks share the same sparsity pattern. In task n, the learner faces a stochastic linear bandit problem for T steps. The linear bandit problem in task n is specified with decision space $\mathcal{D} \subset \mathbb{R}^d$ and a mean-reward function that maps any action $a \in \mathcal{D}$ to $r_n(a) = a^\top \theta_n$ and is parameterized by vector $\theta_n \in \mathbb{R}^d$. We assume that θ_n is S-sparse, meaning that $\|\theta_n\|_0 \le S$. Let $a_n = \text{argmax}_{a\in\mathcal{D}}\, a^\top \theta_n$ and $A_{n,t}$ be the action in time t of task n. The regret is

$$\text{Regret}_N = T\sum_{n=1}^{N} r_n(a_n) - \sum_{n=1}^{N}\sum_{t=1}^{T} r_n(A_{n,t})\,.$$

Viewed as a partial monitoring game, the EXR choice is any pure exploration method aiming to uncover the non-zero elements of the parameter vector, and its regret can be as large as T. The EXT choice executes a standard linear bandit algorithm (e.g. OFUL of Abbasi-Yadkori et al. [1] in the stochastic case) on a subset of the features. If the chosen subset includes the support of θ_n, the regret is $S\sqrt{T}$ and otherwise the regret can be T. If the solution of the pure exploration method is always part of the optimal subset, by the arguments of previous sections, $\text{Regret}_N \le NS\sqrt{T} + \widetilde{O}(T\sqrt{NS\log d})$.

We can use a similar approach in a bandit meta-learning problem where each task is a linear bandit problem and all reward functions lie in a subspace. Related problems are studied by Yang et al. [48] and Tripuraneni et al. [44].

D Details of the experiments

Recall that r_n denotes the reward vector of task n. Also recall that $\mathbf{C}_{\texttt{info}} = B_{T,K} = \widetilde{O}(\sqrt{KT})$, $\mathbf{C}_{\texttt{hit}} = B_{T,s_n} \le \widetilde{O}(\sqrt{KM})$, and $\mathbf{C}_{\texttt{miss}} = O(T)$.

The optimal subset x^* is an M-subset of $[K]$ chosen uniformly at random. When generating the reward vector of task n, the environment first chooses the optimal arm $a_n^* \in x^*$ using a process that will be explained next. Then we sample the optimal reward value $r_n(a_n^*)$ uniformly at random in the range $[\Delta, 1]$, where $\Delta = \sqrt{K\log(N)/T}$. If the task has b optimal arms, $b-1$ other arms are chosen uniformly at random and their reward is set to $r_n(a_n^*)$. Then, rewards of all other arms are sampled uniformly at random in the range $[0, r_n(a_n^*) - \Delta]$. To choose the optimal arm, we consider different settings as explained next.

In the *stochastic* setting, for each task n, we sample the optimal arm location uniformly $a_n^* \sim x^*$.

The *adversarial* setting with a non-oblivious adversary is applicable when the learner maintains a set of discovered arms as in the `G-BASS` algorithm. Here, with probability q_n (defined in Section 3.1), the adversary chooses a new optimal arm so that $a_n^* \in x^* \setminus x_n$, where x_n is the set of discovered arms by the learner up to task n. Otherwise, the next optimal arm is chosen uniformly at random in

x_n. The exploration probabilities are given by

$$p_n = \frac{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}}}{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}} + \sqrt{2(\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}})(\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}})(N-n-1)}}$$
$$q_n = \frac{\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}}}{\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}} + \sqrt{2(\mathbf{C}_{\texttt{info}} - \mathbf{C}_{\texttt{hit}})(\mathbf{C}_{\texttt{miss}} - \mathbf{C}_{\texttt{hit}})(N-n-1)}}.$$

The oblivious adversary is applicable against any learner even if the learner does not maintain a set of discovered arms. Here the adversary simulates an imaginary G-BASS algorithm, and samples new optimal arms and generates the reward sequence with respect to this imaginary learner.

The baseline algorithms include:

- MOSS is an efficient MAB strategy [2]. Optimal MOSS is MOSS played on the set x^*.
- G-BASS: This is Algorithm 1, a greedy algorithm with a Phased Elimination procedure as the exploration subroutine.
- G-BASS^u: This is the special version of G-BASS under the assumption that each task has a unique optimal arm. It is described in Section 3.1. Here, all the actions returned by the exploration subroutine are added to the learned set until we have M learned actions.
- OG^o: This is the algorithm of Streeter and Golovin [42]. To make it more competitive in this setting, we scale their exploration probability γ by a factor of 0.008. Also if their meta learner chooses a set of actions with repetitions, the repetitions are replaced by random actions.
- E-BASS: This is the sample efficient but computationally expensive algorithm shown in Algorithm 1. Given its high computational cost, it is compared with other baselines only in smaller problems.

Figure 2 (copy of Figure 1) shows the experimental results under the Unique Identification assumption (Assumption 2). Performance of G-BASS is close to the optimal performance, while OG^o is not competitive with the simple baseline MOSS. We observe a similar performance in Figure 3 which shows the experimental results when each task has two optimal arms.

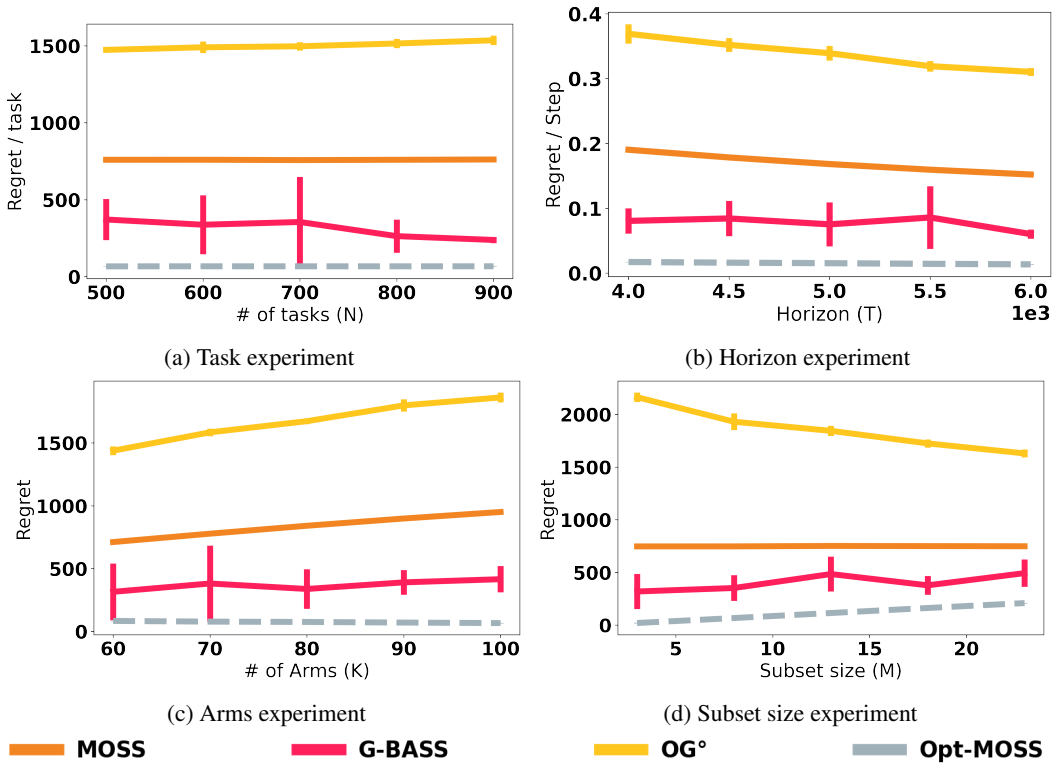
Figures 4 and 5 show experimental results in smaller problems, where E-BASS can also be executed in a reasonable time. Here, E-BASS outperforms other baselines in most settings, while G-BASS and G-BASS^u are also competitive.

The experimental results for the stochastic setting are shown in Figures 5 and 6. The results for the non-oblivious adversarial setting are shown in Figure 9. Note that in this case, the reward sequence is chosen to be nearly the worst case for our proposed learning algorithms. Figures 7 and 8 show experimental results when the algorithm uses a fixed non-adaptive exploration probability $p = \sqrt{\frac{MT}{NB_{T,K}}}$ as in Section 3.2.

In Figures 4-9, as M increases, the performance of the learning methods degrade and can be worse than the simple baseline MOSS. This is because, when M is large, we need a larger number of tasks N to learn the optimal subset. Figure 10 shows that the learning methods still beat the baseline with large enough N.

In Figures 4d and 9d, G-BASS has a lower regret than Optimal MOSS when M is large. This is because, for most of the tasks, G-BASS runs the base bandit algorithm on a small subset that with high probability contains the optimal arm, while Optimal MOSS always runs the base bandit algorithm on a subset of size M. This difference leads to an improved performance for G-BASS specially when M is large.

Figure 11 shows that when the number of arms is large, the length of the first phase of the best-arm identification algorithm (PE) might be larger than horizon T. In that case, G-BASS and G-BASS^u do not learn the meta-structure as the output of the BAI algorithm has poor quality.



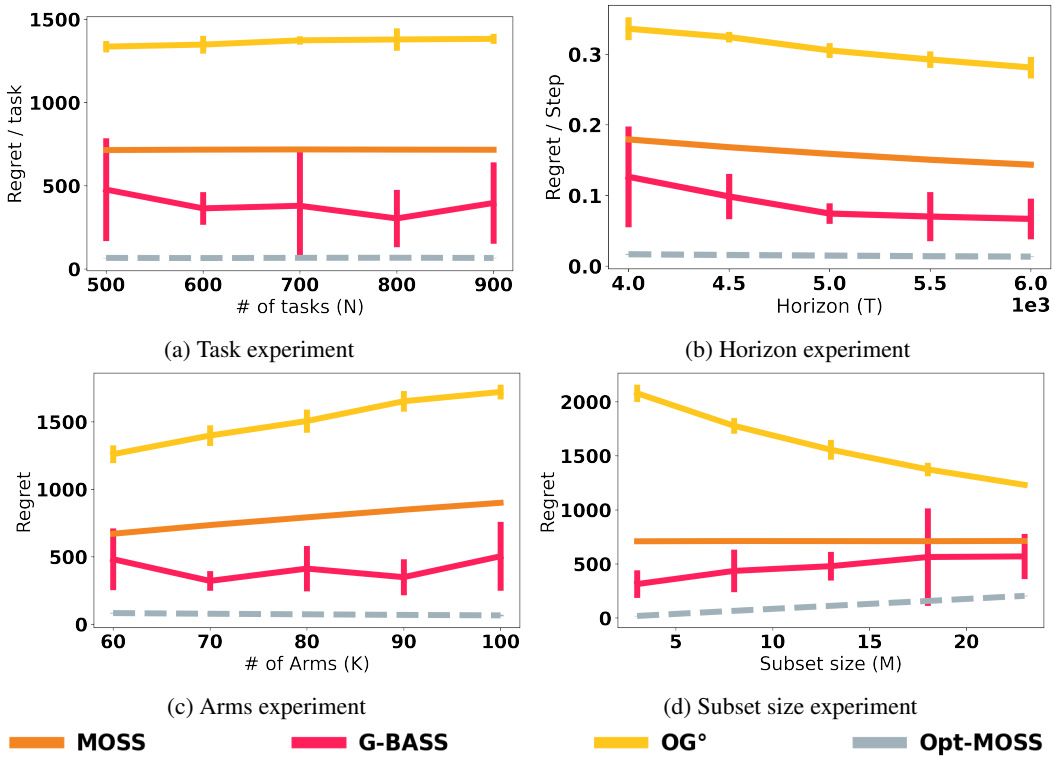
(a) Task experiment

(b) Horizon experiment

(c) Arms experiment

(d) Subset size experiment

Figure 2: Oblivious adversarial setting with one optimal arm per task, and using adaptive p_n. The default setting is $(N, T, K, M) = (500, 4000, 80, 8)$.



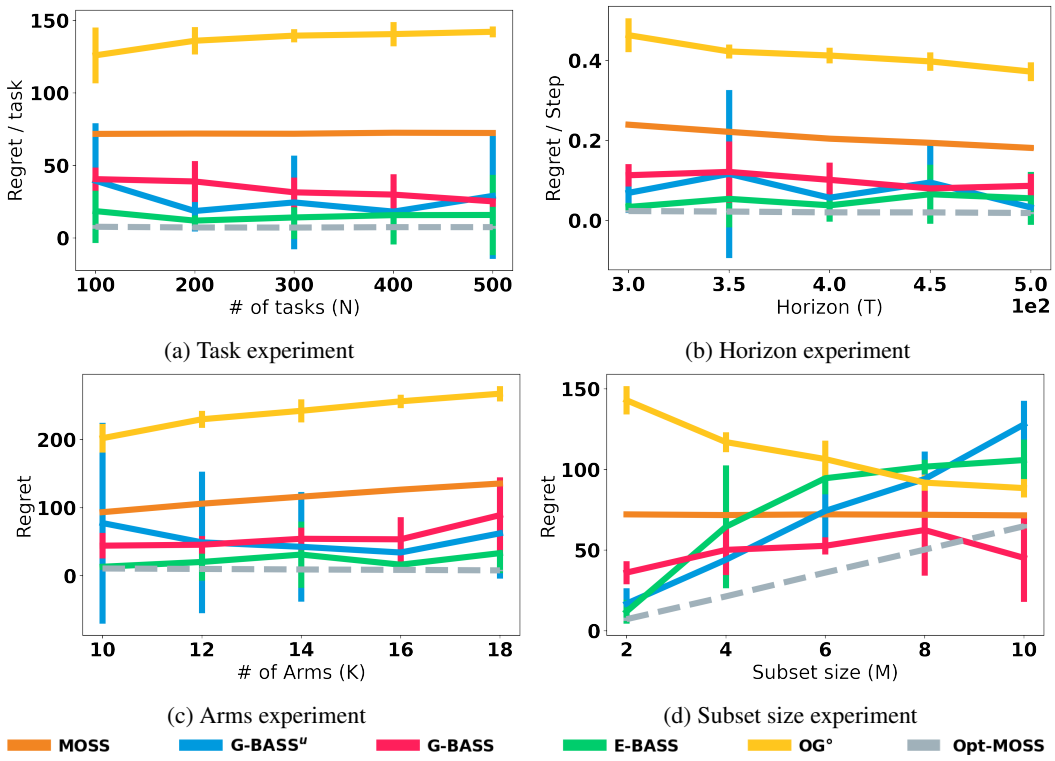
(a) Task experiment

(b) Horizon experiment

(c) Arms experiment

(d) Subset size experiment

Figure 3: Oblivious adversarial setting with two optimal arms per task, and using adaptive p_n. The default setting is $(N, T, K, M) = (500, 4000, 80, 8)$.



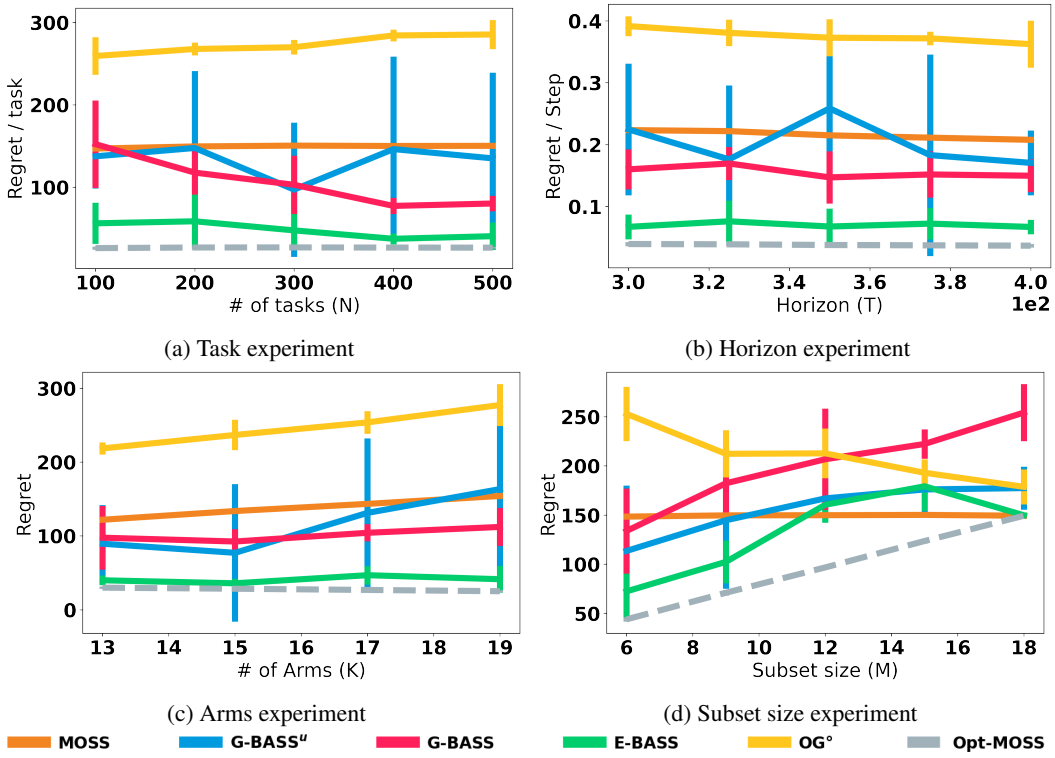
(a) Task experiment

(b) Horizon experiment

(c) Arms experiment

(d) Subset size experiment

Figure 4: Oblivious adversarial setting with one optimal arm per task, and using adaptive p_n. The default setting is $(N, T, K, M) = (200, 300, 11, 2)$.



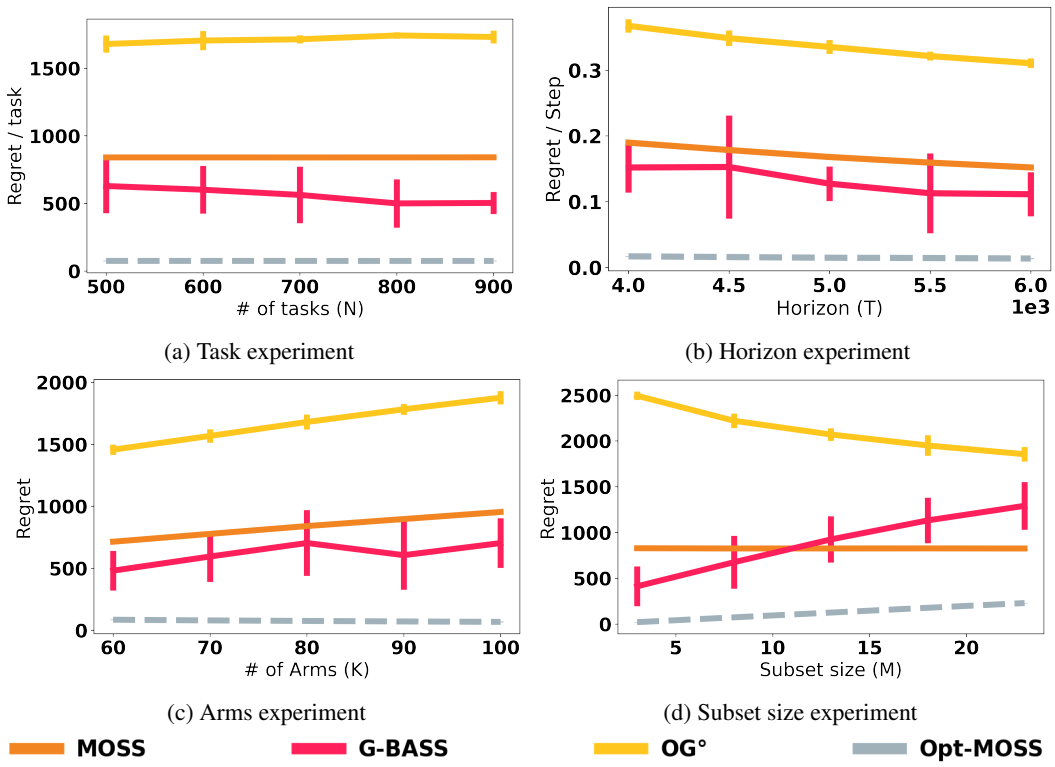
(a) Task experiment

(b) Horizon experiment

(c) Arms experiment

(d) Subset size experiment

Figure 5: The stochastic setting with one optimal arm per task, and using adaptive p_n. The default setting is $(N, T, K, M) = (200, 750, 18, 4)$.



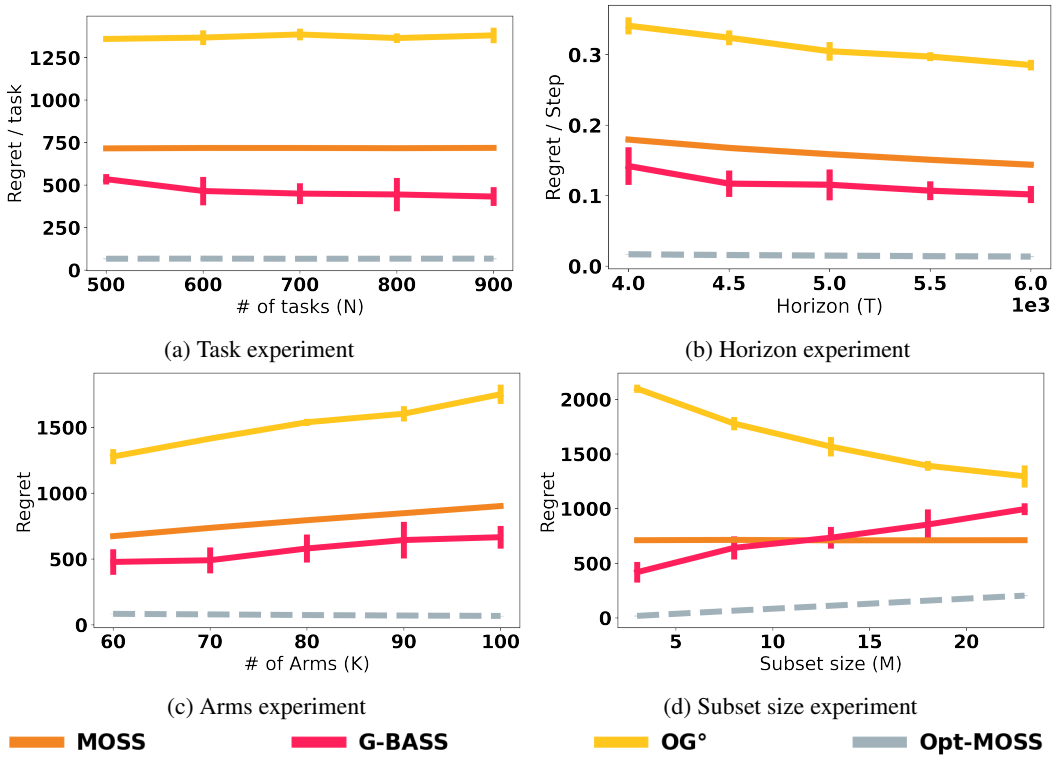
(a) Task experiment

(b) Horizon experiment

(c) Arms experiment

(d) Subset size experiment

Figure 6: The stochastic setting with two optimal arms per task, and using adaptive p_n. The default setting is $(N, T, K, M) = (500, 4000, 80, 8)$.



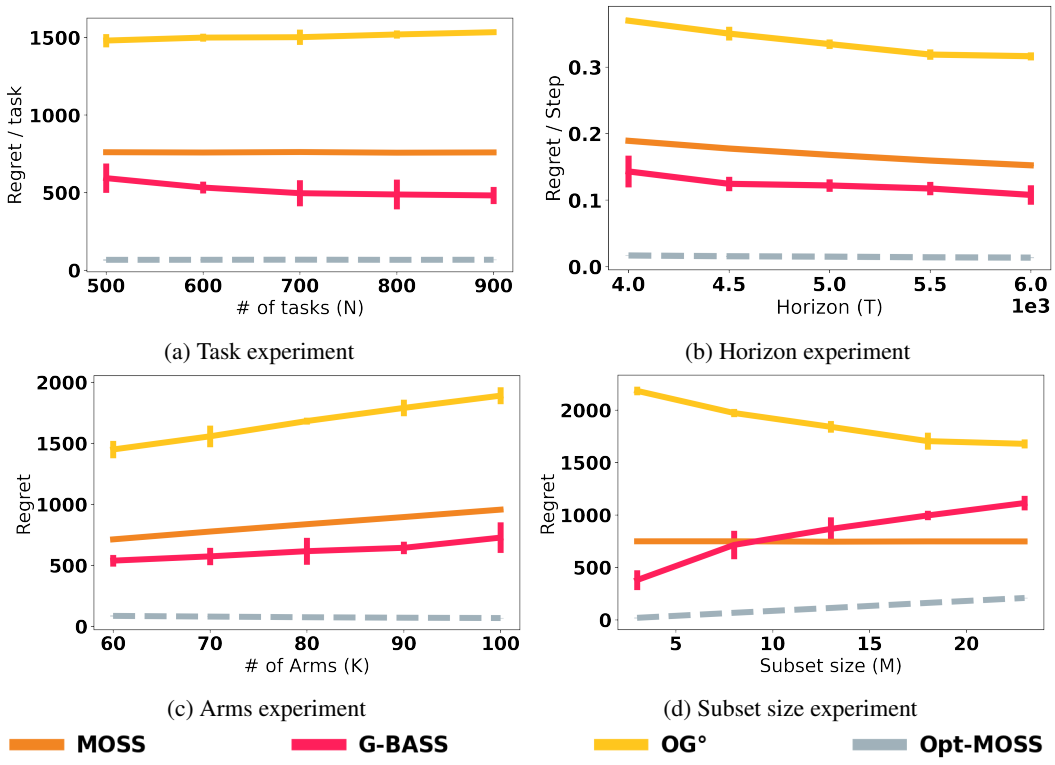
(a) Task experiment

(b) Horizon experiment

(c) Arms experiment

(d) Subset size experiment

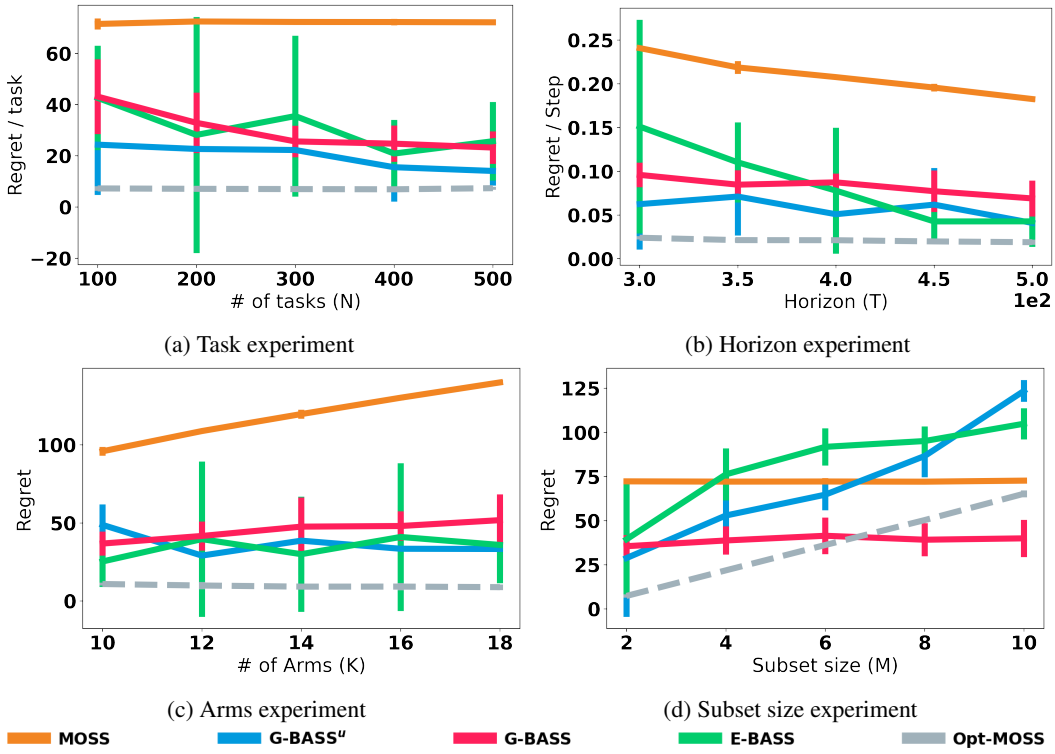
Figure 7: The oblivious adversarial setting with two optimal arms per task, and using fixed p as in Theorem 3.2. The default setting is $(N, T, K, M) = (500, 4000, 80, 8)$.



(a) Task experiment (b) Horizon experiment

(c) Arms experiment (d) Subset size experiment

Figure 8: The stochastic setting with two optimal arms per task, and using fixed p as in Theorem 3.2. The default setting is $(N, T, K, M) = (500, 4000, 80, 8)$.



(a) Task experiment (b) Horizon experiment

(c) Arms experiment (d) Subset size experiment

Figure 9: The non-oblivious adversarial setting with one optimal arm per task, and using adaptive p_n. The default setting is $(N, T, K, M) = (200, 300, 11, 2)$.

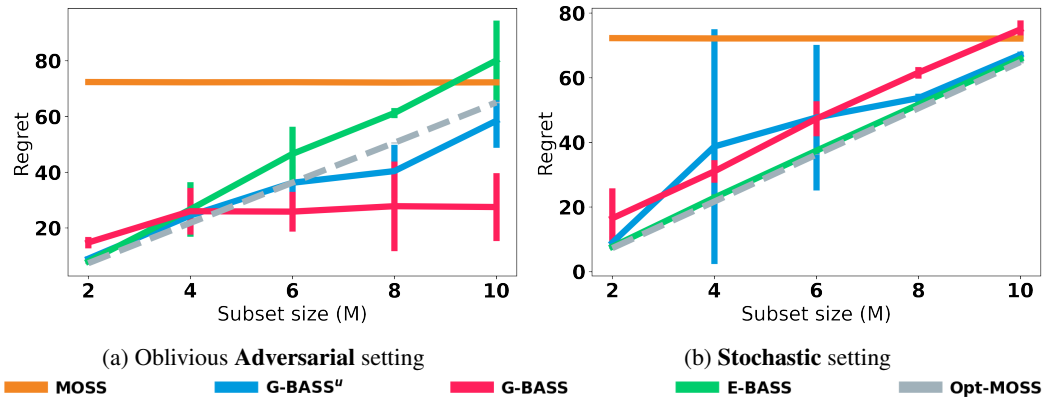


Figure 10: One optimal arm per task and using adaptive p_n. The default setting is $(N, T, K) = (4000, 300, 11)$. When N is large enough, `G-BASS`^{u}, `G-BASS`, and `E-BASS` perform better than the baseline.

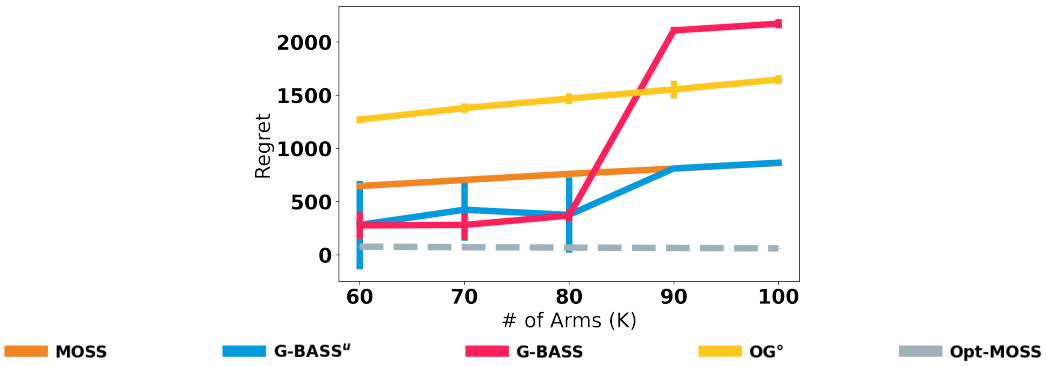


Figure 11: The oblivious adversarial setting with one optimal arm per task, and using adaptive p_n. The default setting is $(N, T, K, M) = (500, 4000, 80, 8)$.

E Gap condition

In this section, we show that the conditions of Section 3 are satisfied. In particular, we show that under a gap condition, the Phased Elimination algorithm executed on the full action set returns the optimal arm with high probability.

Assume all bandit tasks satisfy a gap condition: there exists $\Delta \in (0, 1)$ such that for any $n \in [N]$, if a_n is the optimal arm and a' is any other arm, then $r_n(a_n) - r_n(a') \geq \Delta$. Under this condition, the Phased Elimination of [3] simultaneously satisfies $O(\sqrt{KT})$ cumulative regret and a high probability best-arm identification guarantee.

Given a sequence of integers $(m_i)_{i \geq 1}$, the Phased Elimination algorithm proceeds as follows: in round i, each arm in the remaining set A_i is chosen m_i times. Let $\widehat{\mu}_{a,i}$ be the average reward for arm a using data from phase i. Then all arms such that $\widehat{\mu}_{a,i} + 2^{-i} < \max_{a' \in A_i} \widehat{\mu}_{a',i}$ are eliminated and the remaining arms are in the next set A_{i+1}.

Let arm 1 be the optimal arm. It is known that for any phase i,

$$\mathbf{P}(1 \notin A_{i+1}, 1 \in A_i) \leq K e^{-m_i 2^{-2i}/4} .$$

Further, if $\Delta_a \geq 2^{-i}$, then

$$\mathbf{P}(a \in A_{i+1}, 1 \in A_i, a \in A_i) \leq e^{-m_i (\Delta_a - 2^{-i})^2/4} .$$

Choose $m_i = 4(2^{2i})\log(B)$. We bound the probability that the optimal arm is eliminated in the first I rounds,

$$\mathbf{P}(\exists i \in [I] \,:\, 1 \notin A_i) = \sum_{i=1}^{I} \mathbf{P}(1 \in A_{i-1}, 1 \notin A_i) \leq K \sum_{i=1}^{I} e^{-m_i 2^{-2i}/4} \leq \frac{KI}{B} \,.$$

Let $I_a = \min\{i \geq 1 \,:\, 2^{-i} \leq \Delta_a/2\}$. We bound the probability that a suboptimal arm a remains in phase $I_a + 1$,

$$\begin{aligned}
\mathbf{P}(a \in A_{I_a+1}) &= \mathbf{P}(a \in A_{I_a+1}, a \in A_{I_a}) \\
&= \mathbf{P}(a \in A_{I_a+1}, a \in A_{I_a}, 1 \in A_{I_a}) + \mathbf{P}(a \in A_{I_a+1}, a \in A_{I_a}, 1 \notin A_{I_a}) \\
&\leq \mathbf{P}(a \in A_{I_a+1}, a \in A_{I_a}, 1 \in A_{I_a}) + \mathbf{P}(1 \notin A_{I_a}) \\
&\leq \frac{KI_a + 1}{B} \,.
\end{aligned}$$

The last inequality holds because $\Delta_a \geq 2^{-(I_a-1)} \geq 2^{-I_a}$ and $m_{I_a}(\Delta_a - 2^{-I_a})^2/4 \leq \log B$. We have $I_a = \log(2/\Delta_a) \leq \log(2/\Delta)$. Therefore, $\mathbf{P}(a \in A_{I_a+1}) \leq (1 + K\log(2/\Delta))/B$ and $\mathbf{P}(\exists i \in [I_a] \,:\, 1 \notin A_i) \leq K\log(2/\Delta)/B$. The probability of failure in a bandit task is bounded by $\delta = (1 + 2K\log(2/\Delta))/B$.

In summary and with the choice of $B = N$, if

$$T_a \geq \sum_{i=1}^{I_a} m_i = (4\log B) \sum_{i=1}^{I_a} 2^{2i} = \frac{64\log N}{3\Delta_a^2} \,,$$

the algorithm outputs the optimal arm with probability at least $\delta = (1 + 2K\log(2/\Delta))/N^2$. The cumulative regret of the algorithm is the same as that of the UCB algorithm up to constant terms:

$$R_T \leq \sum_a T_a \Delta_a \leq \frac{64K\log N}{3\Delta_a} \,.$$