A Novel Confidence-Based Algorithm for Structured Bandits

Andrea Tirinzoni
Politecnico di Milano

Alessandro Lazaric
Facebook AI Research

Marcello Restelli
Politecnico di Milano

Abstract

We study finite-armed stochastic bandits where the rewards of each arm might be correlated to those of other arms. We introduce a novel phased algorithm that exploits the given structure to build confidence sets over the parameters of the true bandit problem and rapidly discard all sub-optimal arms. In particular, unlike standard bandit algorithms with no structure, we show that the number of times a suboptimal arm is selected may actually be reduced thanks to the information collected by pulling other arms. Furthermore, we show that, in some structures, the regret of an anytime extension of our algorithm is uniformly bounded over time. For these constant-regret structures, we also derive a matching lower bound. Finally, we demonstrate numerically that our approach better exploits certain structures than existing methods.

1 Introduction

The widely studied multi-armed bandit (MAB) (Lai and Robbins, 1985; Bubeck and Cesa-Bianchi, 2012) problem is one of the simplest sequential decision-making settings in which a learner faces the exploration-exploitation dilemma. At each time t, the learner chooses an *arm* I_t from a finite set $\mathcal{A}$ and receives a random *reward* X_t whose unknown distribution depends on the chosen arm. The goal is to maximize the cumulative reward (or, equivalently, to minimize the regret w.r.t. the best arm) over a horizon n, which requires the agent to trade off between *exploring* arms to understand their uncertain outcomes and *exploiting* those that have performed best in the past.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

The classic MAB problem, in which the rewards of the different arms are uncorrelated, is now theoretically well understood. In their seminal paper, Lai and Robbins (1985) provided the first asymptotic problem-dependent lower bound on the regret. Several simple yet near-optimal strategies have then been proposed, such as UCB1 (Auer et al., 2002), Thompson Sampling (TS, Thompson, 1933), and KL-UCB (Garivier and Cappé, 2011). However, the assumption that the arms are uncorrelated might be too general. In many applications, such as recommender systems or health-care, arms exhibit known structural properties that bandit algorithms could exploit to significantly speed-up the learning process.[1]

Several specific structures have been addressed in the literature. Linear bandits are a well-known example, in which the mean reward of each arm is a linear function of some unknown parameter. Several algorithms have been proposed for these settings, such as extensions of UCB (Abbasi-Yadkori et al., 2011) and TS (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017). However, these approaches, mostly based on the optimism in the face of uncertainty (OFU) principle, have been proved not asymptotically optimal (Lattimore and Szepesvari, 2017). Examples of other specific structures include combinatorial bandits (Cesa-Bianchi and Lugosi, 2012), Lipschitz bandits (Magureanu et al., 2014), ranking bandits (Combes et al., 2015), unimodal bandits (Yu and Mannor, 2011), etc.

Recently, there has been a growing interest in designing bandit strategies to exploit general structures, where the learner is provided with a subset of all possible bandit problems containing the (unknown) problem she has to face. The structured UCB algorithm, proposed almost-simultaneously by Lattimore and Munos (2014) and Azar et al. (2013), applies the OFU principle to general structures. Atan et al. (2018) proposed a greedy algorithm for the special case where all arms are informative, while Wang et al. (2018) extended these settings to consider correlations

[1] In recommender systems, it is often possible to cluster users in a few *types* based on their preferences. Once the *type* of user is known, the value of each item is fixed.

only within certain groups of arms and independence among them. Gupta et al. (2018) generalized UCB and TS to exploit the structure and quickly identify sub-optimal arms. One of the interesting findings of these works is that, in some structures, constant regret (i.e., independent of n) is possible. In the remainder, we shall call these strategies *confidence-based* since they explicitly maintain the uncertainties about the true bandit and use these to trade-off exploration/exploitation. Although conceptually simple, confidence-based strategies are typically hard to design and analyze in a fully structure-aware manner. In fact, in structured problems, pulling an arm provides not only a sample of its mean, but also information about the bandit problem itself through the knowledge of the overall structure. In turn, information about the problem itself potentially allow to refine the estimates of the means of *all* arms. Combes et al. (2017) made a significant step in exploiting this interplay between arms and bandit problems in the very definition of the algorithm itself. The authors derived a structure-aware lower bound characterizing the optimal pull counts as the solution to an optimization problem. Their algorithm, OSSB, approximates this solution and achieves asymptotic optimality for any general structure. However, since the lower bound depends on the true (unknown) bandit at hand, this approach requires to force some exploration to guarantee a sufficiently accurate solution. For this reason, we shall call this kind of strategy *forced-exploration*. Compared to confidence-based ones, it can be intractable in many structures and it remains an open question how well it performs in finite time.

In this paper, we focus on the widely-applied confidence-based strategies for structured bandits. Our contributions are as follows. **1)** We propose an algorithm running through phases. At the beginning of each phase, the set of bandit models compatible with the confidence intervals computed so far is built and the corresponding optimal arms are repeatedly pulled in a round-robin fashion, until the end of the phase. For this strategy, we prove an upper bound on the expected regret that, compared to existing bounds, better shows the potential benefits of exploiting the structure. The key finding is that the number of pulls to a sub-optimal arm i can be significantly reduced by exploiting the information obtained while pulling other arms, and notably the arm that is most informative for this purpose, i.e., the arm for which the mean of the true bandit differs the most from that of any other bandit in which arm i is optimal. This is in contrast to existing methods, which rely exclusively on the samples obtained from arm i to identify its suboptimality (a property that is true for the unstructured settings). **2)** Since our algorithm requires to know the horizon n, we design a practical anytime extension for which, under the same assumptions as in (Lattimore and Munos, 2014), we derive a constant-regret bound with a better scaling in the relevant structure-dependent quantities. **3)** For certain structures that satisfy the aforementioned assumption, we also derive a matching lower bound that shows the optimality of our algorithm in the constant-regret regime. **4)** We report numerical simulations in some simple illustrative structures that confirm our theoretical findings.

2 Preliminaries

We follow similar notation and notions to formalize MAB with structure as in (Agrawal et al., 1988; Graves and Lai, 1997; Burnetas and Katehakis, 1996; Azar et al., 2013; Lattimore and Munos, 2014; Combes et al., 2017). We denote by Θ^{all} the collection of all bandit problems θ with a set of arms $\mathcal{A}$ and whose reward distributions $\{\nu_i\}_{i\in\mathcal{A}}$ are bounded in $[0,1]$[2]. We refer to each $\theta \in \Theta^{all}$ as a *bandit (problem)*, or *model*. We denote by $\mu_i(\theta)$ the mean reward of arm i in model θ and let $\mu^*(\theta) := \max_{i\in\mathcal{A}} \mu_i(\theta)$. For the sake of readability, we assume that the corresponding optimal arm, $i^*(\theta) := \operatorname{argmax}_{i\in\mathcal{A}} \mu_i(\theta)$, is unique for all models. The sub-optimality gap of arm $i \in \mathcal{A}$ is $\Delta_i(\theta) := \mu^*(\theta) - \mu_i(\theta)$, while the model gap w.r.t. $\theta' \in \Theta^{all}$ is $\Gamma_i(\theta,\theta') := |\mu_i(\theta) - \mu_i(\theta')|$. It is known that the gaps Δ characterize the complexity of a bandit problem in the unstructured case. As we shall see, the model gaps Γ play the analogous role in structured problems. A *structure* $\Theta \subseteq \Theta^{all}$ is a subset of possible models. For instance, a linear structure is a set of models whose mean rewards can be written as a linear combination of given features. We denote by $\mathcal{A}^*(\Theta)$, abbreviated $\mathcal{A}^*$ when Θ is clear from context, the set of arms that are optimal for at least one model in Θ, while Θ_i^* is the set of models in which arm i is optimal.

Let $\theta^* \in \Theta^{all}$ be the *true* model and $\Omega := \{\Theta' \subseteq \Theta^{all} \mid \theta^* \in \Theta'\}$. A (structured) bandit algorithm π receives as input a structure $\Theta \in \Omega$ and defines a strategy for choosing the arm I_t given the history $H_{t-1} = (I_1, X_1, \ldots, I_{t-1}, X_{t-1})$[3]. Our performance measure is the expected regret after n steps,

$$R_n^\pi(\theta^*, \Theta) := n\mu^*(\theta^*) - \mathbb{E}_{\pi,\theta^*}\left[\sum_{t=1}^n \mu_{I_t}(\theta^*)\right].$$

Note that the regret depends on Θ through the strategy π. In the remaining, whenever θ is dropped from a model-dependent quantity, we implicitly refer to θ^*.

[2] As usual, this assumption can be relaxed to sub-Gaussian noise with no additional complications.

[3] Whenever π receives as input Θ^{all}, it reduces to the standard MAB case.

Structured UCB Structured UCB (SUCB)[4] is a natural extension of the OFU principle to general structures and it reduces to UCB whenever the structure Θ provided as input is the set of all possible bandit problems (i.e., Θ^{all}). At each step t, the algorithm builds a confidence set $\tilde{\Theta}_t \subseteq \Theta$ containing all the models compatible with the confidence intervals built for each arm and it pulls the optimistic arm $I_t = \text{argmax}_{i \in \mathcal{A}} \sup_{\theta \in \tilde{\Theta}_t} \mu_i(\theta)$. While taking the optimistic arm ensures that "good" arms are selected, refining the confidence set $\tilde{\Theta}_t$ allows to exploit the structure to possibly discard arms more rapidly. Lattimore and Munos (2014) derived the same upper bound to the regret as the one of UCB without making any assumption on set Θ. On the other hand, Azar et al. (2013) derived a more structure-aware bound, but only for finite Θ. The next theorem combines the best of these analyses (see proof in App. B). We first introduce two quantities that conveniently characterize the number of samples needed to distinguish between models. For any $\Theta' \in \Omega$ and $\mathcal{A}' \subseteq \mathcal{A}$, we define:

$$\Psi(\Theta', \mathcal{A}') := \inf_{\theta \in \Theta'} \max_{j \in \mathcal{A}'} \Gamma_j^2(\theta, \theta^*), \tag{1}$$

$$\psi(\Theta', \mathcal{A}') := \operatorname*{arginf}_{\theta \in \Theta'} \max_{j \in \mathcal{A}'} \Gamma_j^2(\theta, \theta^*). \tag{2}$$

It is known that the number of pulls to an arm i that are sufficient to distinguish between θ^* and any θ is bounded as $\mathcal{O}(1/\Gamma_i^2(\theta, \theta^*))$ with high-probability (Azar et al., 2013). Then, we can interpret $\Psi(\Theta', \mathcal{A}')$ as proportional to the inverse number of pulls required from the *most effective* arm in $\mathcal{A}'$ to distinguish θ^* from the model $\psi(\Theta', \mathcal{A}')$, i.e., the bandit problem in Θ' that is most similar to θ^* in terms of model gaps. For this reason, we refer to $\psi(\Theta', \mathcal{A}')$ as the *hardest model* in Θ' using arms in $\mathcal{A}'$. Finally, we define the following sets of optimistic models w.r.t. θ^*: $\Theta^+ := \{\theta \in \Theta : \mu^*(\theta) > \mu^*(\theta^*)\}$ and $\Theta_i^+ := \{\theta \in \Theta^+ : i^*(\theta) = i\}$.

Theorem 1. *There exist constants $c, c' > 0$ such that for any model $\theta^* \in \Theta^{all}$ and any structure $\Theta \in \Omega$, the expected regret at time n of the SUCB algorithm (Lattimore and Munos, 2014) is upper-bounded as*

$$R_n^{SUCB}(\theta^*, \Theta) \leq \sum_{i \in \mathcal{A}^* \setminus \{i^*\}} \frac{c \Delta_i(\theta^*) \log n}{\Psi(\Theta_i^+, \{i\})} + c'.$$

This result shows that SUCB is able to leverage the knowledge of Θ to improve over UCB, which relies only on Θ^{all}. First, the summation is limited to arms that are optimal in at least one model in Θ. Second, the number of pulls of a sub-optimal arm i depends on the model gap $\Gamma_i(\theta_i^+, \theta^*)$ w.r.t. the *hardest model* $\theta_i^+ = \psi(\Theta_i^+, \{i\})$. This measures the number of pulls necessary to distinguish θ_i^+ from θ^* by pulling i. This gap can be much larger than the sub-optimality gap $\Delta_i(\theta^*)$ which appears in unstructured settings (e.g., UCB), thus significantly reducing the final regret.

While UCB-based algorithms are proved to be optimal (i.e., they match the asymptotic lower bound of Lai and Robbins (1985)), evaluating the optimality of Thm. 1 is less obvious. We need to first introduce a specific type of structures. We say that Θ is a worst-case structure if it belongs to the set

$$\Omega^{\text{wc}} := \left\{\Theta \in \Omega \mid \forall i \neq i^* : \Psi(\Theta_i^+, \{i\}) = \Psi(\bar{\Theta}_i^+, \{i\})\right\},$$

where $\bar{\Theta}_i^+ := \{\theta \in \Theta_i^+ | \max_{j \neq i} \Gamma_j(\theta, \theta^*) = 0\}$ is the subset of optimistic models that are indistinguishable from θ^* except in their optimal arm. Thus, a worst-case structure is such that the hardest optimistic models cannot be distinguished from θ^* except in their optimal arm. Note that $\Theta^{all} \in \Omega^{\text{wc}}$. An asymptotic lower bound for these structures has already been provided by Burnetas and Katehakis (1996). We state here the version for Gaussian bandits with fixed variance equal to 1 to facilitate comparison with the upper-bounds.

Theorem 2 (Burnetas and Katehakis (1996))**.** *For any $\Theta \subseteq \Omega^{wc}$ and uniformly convergent strategy π,*

$$\liminf_{n \to \infty} \frac{R_n^\pi(\theta^*, \Theta)}{\log n} \geq \sum_{i \in \mathcal{A}^* \setminus \{i^*\}} \frac{\Delta_i(\theta^*)}{\Psi(\Theta_i^+, \{i\})}.$$

We refer the reader to (Garivier et al., 2018) for a simple proof and the definition of uniformly convergent strategies. The immediate consequence of Theorem 2 is that SUCB is asymptotically order-optimal for all worst-case structures.

3 Structured Arm Elimination

Our structured arm elimination (SAE) strategy (Algorithm 1) is a phased algorithm inspired by Improved UCB (Auer and Ortner, 2010). In each phase h, the algorithm keeps a confidence set containing the models such that the mean of each arm i does not deviate too much from the empirical one $\hat{\mu}_{i,h-1}$ according to its number of pulls $T_i(h-1)$, both computed at the end of the previous phase. Then, all active arms (i.e., those that are optimal for at least one of the models in the confidence set) are played until a well-chosen pull count is reached. Such count is computed to ensure that all models that are sufficiently distant from the target θ^* (according to an exponentially-decaying removal threshold $\tilde{\Gamma}_h$) are discarded from the confidence set. Once all the models in which a certain arm $i \in \mathcal{A}$

[4]The algorithm was originally called UCB-S by Lattimore and Munos (2014) and mUCB by Azar et al. (2013).

Algorithm 1 Structured Arm Elimnation (SAE)

Require: Set of models Θ, horizon n, scalars $\alpha > 0, \beta \geq 1$

Initialization:
$\tilde{\Theta}_0 \leftarrow \Theta$ (confidence set)
$\tilde{A}_0 \leftarrow \mathcal{A}^*(\Theta)$ (set of active arms)
$\tilde{\Gamma}_0 \leftarrow 1$ (removal threshold)
Foreach phase $h = 0, 1, \ldots$ **do**
Play all active arms in a round-robin fashion until $\left\lceil \frac{\alpha \log n}{\tilde{\Gamma}_h^2} \left(1 + \frac{1}{\beta}\right)^2 \right\rceil$ pulls are reached for all $i \in \tilde{A}_h$
Update confidence set:
$\tilde{\Theta}_{h+1} \leftarrow \left\{ \theta \in \Theta \mid \forall i \in \mathcal{A} : |\hat{\mu}_{i,h} - \mu_i(\theta)| < \sqrt{\frac{\alpha \log n}{T_i(h)}} \right\}$[5]
Update set of active arms: $\tilde{A}_{h+1} = \mathcal{A}^*(\tilde{\Theta}_{h+1}) \cap \tilde{A}_h$
Decrease removal threshold: $\tilde{\Gamma}_{h+1} \leftarrow \frac{\tilde{\Gamma}_h}{2}$
End

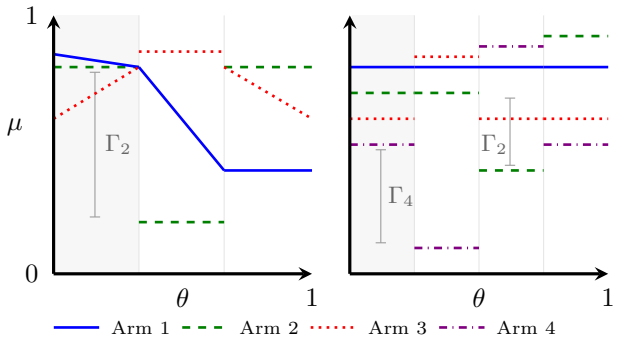

Figure 1: Two structures in which SUCB and SAE significantly differ. The true model is any in the shaded region. *(left)* SUCB never pulls an informative arm. *(right)* SUCB discards an informative arm too early.

is optimal have been eliminated, i is labeled as inactive and no longer pulled. Algorithm 1 can be applied to any set of models (not only finite ones) as far as we can determine the set of optimal arms at each step. This is an optimization problem that can be solved efficiently for, e.g., linear, piecewise-linear, and convex structures, while it becomes intractable in general.

Note that SAE is not an optimistic algorithm since it might pull arms that are never optimistic w.r.t. θ^*. This property is due to the phased nature of the algorithm, such that no *optimistic bias* in selecting the active arms is used, unlike in SUCB. While in unstructured problems SUCB and SAE reduce to UCB and improved UCB, respectively, and have similar regret guarantees (i.e., each arm is pulled roughly the same amount of times in the two algorithms), in structured problems they may behave very differently, as we shall see in the next examples.

3.1 Examples

Figure 1 presents two simple structures in which SUCB and SAE significantly differ. The model set is divided in different regions. Since all bandits in the same region have, for the purpose of our discussion, the same properties, we call θ_1 any model in the first part, θ_2 any model in the second, and so on. Note that the following comments hold for an ideal realization in which certain high-probability events occur.

In the structure of Figure 1*(left)*, arm 2 is never optimistic since its mean is always below the value of the optimal arm $\mu_1(\theta_1)$. Therefore, SUCB never pulls it and needs only to discard the optimistic arm 3. This, in turn, takes $\mathcal{O}(1/\Gamma_3^2(\theta_1, \theta_2))$ pulls of such arm, which can be rather large. Since SAE pulls also arm 2, the large gap $\Gamma_2(\theta_1, \theta_2)$ (Γ_2 in the figure) allows to discard arm 3 much sooner. From the definition of the algorithm, SAE also needs to discard arm 2. Once again, this can be done quickly due to the large gap $\Gamma_1(\theta_1, \theta_3)$ and the fact that the optimal arm 1 is always pulled.

In the structure of Figure 1*(right)*, the optimistic bias makes SUCB pull the arms starting from the one with the highest value, arm 2, downwards to the optimal one, arm 1. Since the gap $\Gamma_2(\theta_1, \theta_3)$ (Γ_2 in the figure) is larger than $\Gamma_2(\theta_1, \theta_4)$, SUCB implicitly discards θ_3, and so arm 4, before arm 2. Thus, once both these arms have been eliminated, the algorithm takes $\mathcal{O}(1/\Gamma_3^2(\theta_1, \theta_2))$ pulls of arm 3 to discard the arm itself. By simultaneously pulling all four arms, SAE discards arm 3 first using the pulls of arm 4 (the one prematurely discarded by SUCB) due to the large gap $\Gamma_4(\theta_1, \theta_2)$ (Γ_4 in the figure). Finally, the deletion of the remaining two sub-optimal arms occurs with the same number of pulls as SUCB, and it can be verified that the overall regret is much smaller.

3.2 Regret Analysis

In order to upper bound the regret of Alg. 1, we need to characterize the arms pulled in each phase, which are specified by the sets of active arms $\left\{\tilde{A}_h\right\}_h$. Since these sets are random quantities, we cannot study them directly. Instead, we introduce a deterministic sequence of active arm sets $\{\mathcal{A}_h\}_h$ that effectively works as a proxy for $\left\{\tilde{A}_h\right\}_h$ and, under certain high-probability events, allows us to define how many samples are needed for arms to be discarded. We now provide intuitions (made formal in the proof of the regret bound) on how such sequence is built. Clearly, we have $\mathcal{A}_0 = \tilde{A}_0 = \mathcal{A}^*(\Theta)$ by definition. Since all arms in $\mathcal{A}_0$ are pulled in $h = 0$, and recalling the meaning of Ψ (Equation 1), our well-chosen pull counts are *sufficient* to prove that all arms i such that $\Psi(\Theta_i^*, \mathcal{A}_0) \geq \tilde{\Gamma}_0^2$

[5] We implicitly assume this condition to hold for arms that have never been pulled before.

are discarded. Let us call the set of these discarded arms $\bar{\mathcal{A}}_0$ and apply this reasoning inductively by setting $\mathcal{A}_1 = \mathcal{A}_0 \setminus \bar{\mathcal{A}}_0$. Unfortunately, it is general not possible to conclude that $\mathcal{A}_1 = \tilde{\mathcal{A}}_1$ since other arms might be discarded. Therefore, we build an additional set $\underline{\mathcal{A}}_h$ of those arms that are guaranteed to be active in phase h. The main intuition is that, if we can prove that certain arms are still active, we can also show that the algorithm uses their *information* (i.e., the model-gaps) to discard certain other arms/models faster. Imagine that an oracle provides us with the set $\underline{\mathcal{A}}_h$. Then, for $h \geq 0$ we have

$$\bar{\mathcal{A}}_h := \left\{ i \in \mathcal{A}_h \mid \tilde{\Gamma}_h \leq \inf_{\theta \in \Theta_i^*} \max_{j \in \underline{\mathcal{A}}_h \cup \{i\}} \Gamma_j(\theta, \theta^*) \right\},$$

with $\mathcal{A}_0 = \mathcal{A}^*(\Theta)$ and $\mathcal{A}_{h+1} = \mathcal{A}_h \setminus \bar{\mathcal{A}}_h$ for $h \geq 1$. Given these sets, we have $\underline{\mathcal{A}}_0 := \mathcal{A}^*(\Theta)$ and

$$\underline{\mathcal{A}}_h := \left\{ i \in \mathcal{A}_h \mid \tilde{\Gamma}_{h-1} > k_\beta \inf_{\theta \in \Theta_i^*} \max_{j \in \mathcal{A}^*(\Theta)} \frac{\Gamma_j(\theta, \theta^*)}{2^{[h - \bar{h}_j - 1]_+}} \right\}$$

for all $h \geq 1$, where $k_\beta := \frac{1}{\beta - 1} \sqrt{(\beta+1)^2 + \frac{1}{\log n}}$ and $\bar{h}_j := \max_{h \in \mathbb{N}^+} \{h \mid j \in \mathcal{A}_h\}$ is the last phase in which arm j is active in our deterministic sequence $\{\mathcal{A}_h\}_h$. This is essentially the set of arms for which the number of pulls to the active arms at the previous phase is below the removal threshold by a *margin* (defined by k_β). Finally, we define the set of arms that are active in the last phase when i is active as $\mathcal{A}_i^* = \underline{\mathcal{A}}_{\bar{h}_i} \cup \{i\}$.

The following theorem is the key result of this paper. It shows that the regret incurred by SAE for arm i is inversely proportional to the maximum model-gap (taken over the set of arms that are active when arm i is discarded) w.r.t. the hardest model in Θ_i^*.

Theorem 3. *Let $\beta \geq 1$, $\alpha = \beta^2$, $n \geq 64$, and $c_\beta := 4(1+\beta^2)$. Then,*

$$R_n^{SAE}(\theta^*, \Theta) \leq \sum_{i \in \mathcal{A}^* \setminus \{i^*\}} \frac{c_\beta \Delta_i(\theta^*) \log n}{\Psi(\Theta_i^*, \mathcal{A}_i^*)} + 2|\mathcal{A}^*(\Theta)|.$$

One of the key novelties, and complications, in the proof (reported in App. C) is that, in order to carry out a fully structure-aware analysis, we do not only care about proving that sub-optimal arms are not pulled after certain phases, but also about guaranteeing that some arms are not discarded too early since their pulls might allow to discard other models/arms. The parameter β plays an important role for this purpose. In particular, k_β controls the sets of arms that, with high probability, are guaranteed to be active at certain phases. For example, for large n, setting $\beta = 3$ yields $k_\beta \simeq 2$, which in turn implies that $\underline{\mathcal{A}}_h$ is the set of arms such that $\tilde{\Gamma}_h > \inf_{\theta \in \Theta_i^*} \max_{j \in \mathcal{A}^*(\Theta)} \frac{\Gamma_j(\theta, \theta^*)}{2^{[h - \bar{h}_j - 1]_+}}$. This is close to saying that all the arms that are not eliminated in phase h are also active in such phase.

3.3 Discussion

First, as a sanity check, we verify that the regret bound of Theorem 3 is never worse than the one of UCB. That is, SAE is never negatively affected by the knowledge of the structure and, whenever applied to unstructured problems, the algorithm is, apart from multiplicative/additive constants, finite-time optimal.

Proposition 1. *The SAE algorithm is always sub-UCB, in the sense that there exist constants $c, c' > 0$ such that its regret satisfies*

$$R_n^{SAE}(\theta^*, \Theta) \leq \sum_{i \in \mathcal{A} \setminus \{i^*\}} \frac{c \log n}{\Delta_i(\theta^*)} + c'.$$

The key property of Thm. 3 is that the regret suffered for discarding a sub-optimal arm i does not necessarily scale with the model gaps of such arm (i.e., $\Psi(\Theta_i^*, \{i\})$) but with those of the most effective arm in $\mathcal{A}_i^*$. Thus, compared to SUCB, in which the elimination of a model $\theta \in \Theta_i^*$ requires $\mathcal{O}(1/\Gamma_i^2(\theta, \theta^*))$ pulls of arm i, SAE needs only $\mathcal{O}(1/\max_{j \in \mathcal{A}_i^*} \Gamma_j^2(\theta, \theta^*))$, which is by definition always smaller. Note that, to be precise, SUCB can potentially eliminate models using the pulls of any arm since the confidence sets are built as in SAE. However, in general, it is not possible to prove the same regret bound since the optimism induces a specific pull order that might prevent the algorithm from choosing the arm with the largest model gap. Obviously, SAE does not know this arm in advance and, therefore, ensures it is pulled by choosing all active arms. However, the additional regret incurred to achieve this property can make the algorithm, in some cases, worse than SUCB. In fact, a key difference is that SUCB stops playing a sub-optimal arm i when all optimistic models in Θ_i^+ are discarded, while SAE needs to eliminate all models in which arm i is optimal (even non-optimistic ones). Therefore, although SAE improves the elimination of all optimistic models, it suffers further regret for discarding non-optimistic ones and, in general, the two algorithms are not comparable. A special case are those structures in which the hardest models for each arm i are in the optimistic set, $\psi(\Theta_i^*, \mathcal{A}_i^*) \in \Theta_i^+$, in which SAE improves over SUCB. These *optimistic* structures are defined as:

$$\Omega^{\text{opt}} := \{\Theta \in \Omega \mid \forall i \neq i^* : \Psi(\Theta_i^+, \mathcal{A}_i^*) = \Psi(\Theta_i^*, \mathcal{A}_i^*)\}.$$

Proposition 2. *If $\Theta \in \Omega^{opt}$, SAE is sub-SUCB, in the sense that its regret can be upper bounded by the one of Theorem 1.*

Since SUCB is order-optimal in Ω^{wc} and SAE is sub-SUCB in Ω^{opt}, Theorem 2 immediately implies that SAE is order optimal in $\Omega^{\text{wc}} \cap \Omega^{\text{opt}}$. Although we are

Algorithm 2 Anytime SAE (ASAE)

Require: Set of models Θ, scalars $\alpha > 0, \beta \geq 1, \eta > 0$

Initialization: $\tilde{n}_0 \leftarrow 2$, $\tilde{\Theta}^{-1} \leftarrow \Theta$
Foreach period $k = 0, 1, \dots$ **do**
Initialize confidence sets: $\tilde{\Theta}_0^k \leftarrow \tilde{\Theta}^{k-1}$, $\tilde{\mathcal{A}}_0^k \leftarrow \mathcal{A}^*(\tilde{\Theta}_0^k)$
Run Algorithm 1 with $n = \tilde{n}_k$, $\tilde{\Theta}_0 = \tilde{\Theta}_0^k$, and $\tilde{\mathcal{A}}_0 = \tilde{\mathcal{A}}_0^k$
Update horizon: $\tilde{n}_{k+1} \leftarrow \tilde{n}_k^{1+\eta}$
End

able to guarantee the optimality in less cases, Proposition 2 ensures that SAE improves over SUCB in a wide variety of structures. Unfortunately, we were not able to prove the optimality of our algorithm in any structure besides the worst-case ones.

4 Anytime SAE and Constant Regret

Algorithm 1 cannot be applied whenever the horizon n is unknown, as the length of each phase explicitly depends on it. This has the additional drawback of preventing constant regret from being achieved since a $\log n$ term naturally appears in the resulting bound. As shown by Lattimore and Munos (2014), there exist structures in which constant regret can be obtained and it would be desirable for our strategy to exploit this fact. We, therefore, propose an anytime extension (Algorithm 2). The idea is once again similar to the one by Auer and Ortner (2010): we split the horizon into different *periods* with exponentially increasing length. Therefore, in Algorithm 2, and throughout this section, we overload our notation by adding a superscript k to denote the period of each period-dependent quantity. The key property is that our approach does not reset in each period (as Auer and Ortner (2010) do) but retains the last confidence sets. Though this makes the proofs more involved, we shall see that it allows us to guarantee a constant regret. One can see the analogy between our non-resetting phased approach and the standard way of handling unknown horizons in online algorithms. In the latter case, we typically replace $\log n$ with $\log t$ in the confidence sets, while here we do the same with $\log \tilde{n}_k$. Then, after proving that certain high-probability events occur at each time/period, we can carry out the proofs without forcing any reset.

Due to the additional complications introduced by the anytime extension (in particular, controlling the sets $\underline{\mathcal{A}}_h$), we were able to prove only a weaker bound than the one in Theorem 3 which, however, retains the same benefits. The proofs are reported in Appendix D.

Theorem 4. *Let $\eta = 1$, $\alpha = 2$, and $\beta = 1$. Then,*

$$R_n^{ASAE}(\theta^*, \Theta) \leq \sum_{i \in \mathcal{A}^* \setminus \{i^*\}} \frac{192 \Delta_i(\theta^*) \log n}{\Psi(\Theta_i^*, \{i, i^*\})} + 6|\mathcal{A}^*(\Theta)|.$$

The new bound has the same form as the one of Algorithm 1, except for the fact that the set of active arms for eliminating each i is reduced to $\{i, i^*\} \subseteq \mathcal{A}_i^*$. Note, however, that the presence of these two arms is enough to prove Proposition 1 and 2.

Remark 1. *Algorithm 2 is sub-UCB and, under the same conditions as in Proposition 2, is also sub-SUCB.*

We now prove a constant-regret bound for Algorithm 2. We need the following assumption from (Lattimore and Munos, 2014), which was proven both necessary and sufficient to achieve constant regret.

Assumption 1 (Informative optimal arm)**.** *The structure Θ satisfies*

$$\Gamma_* := \inf_{\theta \in \Theta \setminus \Theta_{i^*}^*} \Gamma_{i^*}(\theta, \theta^*) > 0.$$

In words, when a model is Γ^*-distant (or less) in arm i^* from θ^*, its optimal arm is still i^*. Therefore, pulling i^* eventually discards all sub-optimal arms. This is fundamental to guarantee that, after the algorithm has pulled i^* a sufficient number of times, no sub-optimal arm can become active again due to the increasing period length (hence we choose i^* forever).

Theorem 5. *Let $\eta = 1$, $\alpha = \frac{5}{2}$, $\beta = 1$, $\bar{t} := \frac{20|\mathcal{A}^*(\Theta)| \log 2}{\Gamma_*^2} + 2|\mathcal{A}^*(\Theta)|$, and suppose Assumption 1 holds. Then,*

$$R_n^{ASAE}(\theta^*, \Theta) \leq \sum_{i \in \mathcal{A}^* \setminus \{i^*\}} \frac{480 \Delta_i(\theta^*) \log \bar{t}}{\Psi(\Theta_i^*, \{i, i^*\})} + 9|\mathcal{A}^*(\Theta)|.$$

This bound improves over the one shown by Lattimore and Munos (2014) for SUCB in its dependence on $\bar{t}$, which can be understood as the time at which the algorithm transitions to the constant regret regime. While Lattimore and Munos (2014) proved $\bar{t} \simeq \mathcal{O}(\max\{1/\Gamma_*^2, 1/\Delta_{\min}^2\})$, here we show that such time does not depend on the minimum gap $\Delta_{\min} = \min_{i:\Delta_i(\theta^*)>0} \Delta_i(\theta^*)$. This is intuitive since, by Assumption 1, $\mathcal{O}(1/\Gamma_*^2)$ pulls of i^* should be enough to identify the optimal arm. Although the analysis of SUCB can be improved by replacing the minimum suboptimality gap with the minimum model gap, it seems that this dependence is tight. As an example, consider a structure in which the optimal arm is very informative ($\Gamma_* \gg 0$) but never optimistic. SUCB will never pull it until all optimistic models are discarded, which requires $\mathcal{O}(1/\Gamma_{\min}^2)$ steps in the worst case. Note that, whenever it is applied to structures satisfying Assumption 1, the bound of Theorem 4 does not show constant regret since the proof uses an implicit worst-case argument (i.e., Assumption 1 is assumed false).

5 Constant-Regret Lower Bound

We have seen that SUCB and SAE are order-optimal for structures in Ω^{wc} and $\Omega^{\text{wc}} \cap \Omega^{\text{opt}}$, respectively. One might wonder whether we can still guarantee optimality in some structures where constant regret is achievable (i.e., when Assumption 1 holds). We answer this question affirmatively by deriving a finite-time lower bound on the expected regret of any 'good' strategy. Note that the problem is non-trivial since, under Assumption 1, one cannot build hard models that differ from the true bandit only in the mean of one arm as in the proof of standard lower-bounds (e.g., Burnetas and Katehakis, 1996). Before stating our result, we specify the class of strategies under consideration. We shall use the following definition due to Garivier et al. (2018), which have been adopted to derive finite-time lower-bounds.

Definition 1 (Super-fast convergence)**.** *A strategy π is super-fast convergent on a set Θ if there exists a constant $c > 0$ such that, for any model $\theta \in \Theta$ and sub-optimal arm $i \in \mathcal{A}$, it satisfies*

$$\mathbb{E}_\theta[T_i(n)] \leq \frac{c \log n}{\Delta_i(\theta)^2}.$$

It is easy to see that UCB, SUCB, and SAE are examples of super-fast convergent strategies. Furthermore, we call the class of structures considered in the lower bound *worst-case constant regret* and define it as

$$\Omega^{\text{cr}} := \{\Theta \in \Omega \mid \forall \theta \in \Theta \setminus \Theta^*_{i^*} : \Gamma_{i^*}(\theta, \theta^*) = \Gamma_* \wedge \Gamma_j(\theta, \theta^*) = 0 \ \forall j \neq i^*(\theta), i^*\}.$$

This can be understood as a generalization of the worst-case structure to make Assumption 1 hold. Due to the challenges in deriving the lower bound for large Γ_*, we also need to assume that $0 < \Gamma^* \leq \mathcal{O}\left(\sqrt{\frac{1}{\sum_{i \neq i^*} \Delta_i^{-2}(\theta^*)}}\right)$, with the precise dependence given in Appendix E. Note that Γ_* is a function of the structure and the dependence was omitted for conciseness. We are now ready to state our result.

Theorem 6. *Let $\Theta \in \Omega^{cr}$ and $n \geq \frac{1}{\Gamma_*^2}$. Then, for sufficiently small Γ^*, the expected regret of any super-fast convergent strategy π can be lower bounded by*

$$R_n^\pi(\theta^*, \Theta) \geq \sum_{i \in \mathcal{A}^* \setminus \{i^*\}} \frac{\Delta_i(\theta^*)}{2\Psi(\Theta_i^*, \{i\})} \log \frac{\Delta^2}{4e^2 c \Gamma_*^2 \log \frac{1}{\Gamma_*^2}},$$

where $\Delta := \inf_{\theta' \in \Theta \setminus \Theta^*_{i^*}} \Delta_{i^*}(\theta')$.

The proof, which combines ideas from Garivier et al. (2018) and Degenne et al. (2018), is reported in Appendix E. Note that the lower bound is positive for sufficiently small Γ_*. Apart from other constants, the dependence on Γ_* matches the upper bound of Theorem 5. However, Theorem 5 seems tighter due to the larger set of arms in Ψ at the denominator. This is not surprising since the lower bound considers only structures with well-chosen hard models. It is easy to prove that, when SAE or SUCB are applied to structures in Ω^{cr}, the two bounds match.

Other lower bounds for constant-regret settings have recently been derived. Bubeck et al. (2013) showed that, for the classic unstructured problems, it is enough to know μ^* and a lower bound on the minimum gap to achieve a constant regret. Garivier et al. (2018) refined this result by showing that the knowledge of μ^* alone actually suffices. Lattimore and Munos (2014) studied several specific structured problems where constant regret is (or is not) possible, providing both lower bounds and algorithms to match them. Finally, we note that the asymptotic lower bound by Combes et al. (2017) is zero when Assumption 1 holds as the regret scaled by $\log n$ correctly vanishes as n grows. Their algorithm reduces to a greedy strategy in this setting which is not necessarily finite-time optimal according to Theorem 6.

6 Numerical Simulations

We perform two different classes of experiments. In the first one, we consider well-chosen structures that allow us to better understand the behavior of all algorithms. In the second one, we randomize the structures to provide a more general comparison. In all experiments, we run SAE and its anytime version (ASAE), SUCB, and UCB on Bernoulli bandits. We also compared to the WAGP algorithm of Atan et al. (2018), which however incurred linear regret in all our experiments (their assumptions never hold in our structures) and, therefore, is omitted from the plots. We use $\alpha = 2$ for all algorithms and $\beta = 1$ for SAE. Each plotted curve is the average of 100 independent runs with 95% Student's t confidence intervals.

Hand-coded Structures We first consider the structure of Figure 1*(left)*. We set $n = 10,000$ and $\eta = 0.1$. The results are shown in Figure 2a. SUCB suffers a large regret for removing models in which arm 3 is optimal. On the other hand, SAE quickly discards these models by pulling arm 2, which, in turn, is eliminated by pulling arm 1. Hence the much lower regret, with the anytime version that performs slightly better. Notice also that Assumption 1 is verified and SAE obtains constant regret. SUCB eventually transitions to constant regret too but needs a longer horizon. Alternatively, we can show an example where SUCB is expected to perform better. We modify the structure of Figure 1*(left)* to make arm 2 non-informative (i.e.,

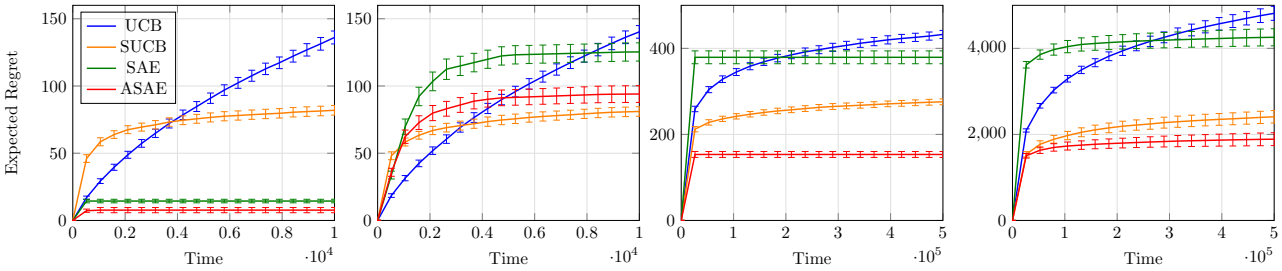

Figure 2: Expected regret in (a) the structure of Figure 1 *(left)*, (b) the same structure with non-informative arm 2, (c) the structure of Figure 1 *(right)*, and (d) randomly-generated structures.

we set its mean to the highest value in the figure for all models) and run the experiment under the same setting. Figure 2b shows that, as expected, SAE suffers from some additional regret for discarding the useless arm and performs worse than SUCB. However, it remains sub-UCB as proved in Section 3.3.

We now consider the structure of Figure 1 *(right)*. We set $n = 500,000$, $\eta = 0.01$, and report the results in Figure 2c. The arm ordering induced by SUCB (from the most optimistic to the optimal one) leads the algorithm to discard arm 4 before even pulling it once. Such arm, however, could be used to quickly discard arm 3, which is what SAE does. Notice that the larger regret of SAE with respect to its anytime counterpart is mainly due to the fact that phased procedures update the confidence sets much less than online approaches. This drawback is alleviated in the anytime version, which reduces the duration of some of these phases and retains good empirical performance.

Randomized Structures We now consider random structures. In each run, we first randomize a set of 100 models with 50 arms by drawing their means from the uniform distribution and we randomly choose the true model among them. Then, we build 50 additional 'hard' models by perturbing a random arm of the true model to become optimal and optimistic, and another random arm to become informative. In particular, the mean of the first random arm is set to $\mu^*(\theta^*) + 0.2\epsilon$, with $\epsilon \sim \mathcal{U}([0,1])$, while the second to 1/10 of the original mean (so that we potentially get a larger model gap). The results are shown in Figure 2d. Most of the regret suffered by SUCB is due to the hard instances we introduced. Some of them are likely to be eliminated by informative arms, but this is not always guaranteed by the SUCB strategy. Both versions of SAE, on the other hand, implicitly exploit these informative arms, with the anytime version outperforming all alternatives. Once again, the original version suffers a high initial regret due to the phased procedure.

7 Discussion

Similarly to most of related literature, our SAE algorithm confirms that simple confidence-based strategies can be designed to exploit general structures, though so far they have been proven optimal only for worst-case structures. Although it only pulls potentially-optimal arms, SAE is not optimistic. The design of non-optimistic algorithms is a key step towards optimality since it is known that OFU-based strategies are not optimal for general structures (Lattimore and Szepesvari, 2017; Combes et al., 2017; Hao et al., 2019). Our regret bounds fully reflect the structure-awareness and their derivation might be of independent interest for analyzing other approaches. Although considering phased strategies is one of our key choices to both obtain the desired algorithmic properties and simplify the proofs, we show empirically that SAE does not suffer from it too much. In particular, it outperforms online strategies in specific structures where informative arms exist that are not always pulled with the OFU principle.

The key open question is how to design confidence-based strategies that are optimal for general structures. The algorithms discussed in this paper have been proven optimal only for certain worst-case structures, while algorithms like OSSB are asymptotically optimal for general structures but require to force exploration to solve an oracle optimization problem. Whether the optimal pull counts of a lower-bound like the one by Combes et al. (2017) can be attained in confidence-based settings and with good finite-time performance remains unknown. We believe that recent advances in the context of pure exploration for bandit problems (Ménard, 2019; Degenne et al., 2019) might provide useful insights into this problem. Furthermore, a finite-time extension of the asymptotic lower bound for general structures, and the corresponding design of finite-time optimal algorithms, is a challenging but interesting research direction.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Abeille, M. and Lazaric, A. (2017). Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184.

Agrawal, R., Teneketzis, D., and Anantharam, V. (1988). Asymptotically efficient adaptive allocation schemes for controlled markov chains: Finite parameter space. In *Proceedings of the 27th IEEE Conference on Decision and Control*, pages 1198–1203. IEEE.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Atan, O., Tekin, C., and van der Schaar, M. (2018). Global bandits. *IEEE transactions on neural networks and learning systems*, 29(12):5798–5811.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Auer, P. and Ortner, R. (2010). Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

Azar, M., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 2220–2228.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Bubeck, S., Perchet, V., and Rigollet, P. (2013). Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134.

Burnetas, A. N. and Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142.

Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422.

Combes, R., Magureanu, S., and Proutiere, A. (2017). Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1763–1771.

Combes, R., Magureanu, S., Proutiere, A., and Laroche, C. (2015). Learning to rank: Regret lower bounds and efficient algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):231–244.

Degenne, R., Garcelon, E., and Perchet, V. (2018). Bandits with side observations: Bounded vs. logarithmic regret. *arXiv preprint arXiv:1807.03558*.

Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems*, pages 14465–14474.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376.

Garivier, A., Ménard, P., and Stoltz, G. (2018). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*.

Graves, T. L. and Lai, T. L. (1997). Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743.

Gupta, S., Joshi, G., and Yağan, O. (2018). Exploiting correlation in finite-armed structured bandits. *arXiv preprint arXiv:1810.08164*.

Hao, B., Lattimore, T., and Szepesvari, C. (2019). Adaptive exploration in linear contextual bandit. *arXiv preprint arXiv:1910.06996*.

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

Lattimore, T. and Munos, R. (2014). Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558.

Lattimore, T. and Szepesvari, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737.

Magureanu, S., Combes, R., and Proutiere, A. (2014). Lipschitz bandits: Regret lower bounds and optimal algorithms. *arXiv preprint arXiv:1405.4758*.

Ménard, P. (2019). Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*.

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

Wang, Z., Zhou, R., and Shen, C. (2018). Regional multi-armed bandits. *arXiv preprint arXiv:1802.07917*.

Yu, J. Y. and Mannor, S. (2011). Unimodal bandits. In *ICML*, pages 41–48. Citeseer.