

Information maximization for generalized multiarmed bandit games

Alex Barbier–Chebbah^{✉,*}, Christian L. Vestergaard[✉], and Jean-Baptiste Masson^{✉,†}

*Institut Pasteur, Université Paris Cité, CNRS UMR 3571, Decision and Bayesian Computation, 75015 Paris, France
and Épiphémie, Inria, Paris, France*



(Received 17 June 2025; accepted 24 November 2025; published 20 January 2026)

Information and free-energy maximization are physics principles that provide general rules for an agent to optimize actions in line with specific goals and policies. These principles can serve as building blocks for designing decision-making policies capable of efficient performance with only partial information. Notably, the information-maximization principle has shown remarkable success in the classical bandit problem and has recently been shown to yield optimal algorithms for Gaussian and sub-Gaussian reward distributions. This article explores a broad extension of physics-based approaches to more complex and structured bandit problems. To this end, we cover three distinct types of bandit problems, where information maximization is adapted and leads to strong performance for Gaussian and Bernoulli reward distributions. Since the main challenge of information maximization lies in avoiding overexploration, we highlight how information is tailored at various levels to mitigate this issue, paving the way for more efficient and robust decision-making strategies.

DOI: [10.1103/wkw4-94hz](https://doi.org/10.1103/wkw4-94hz)

I. INTRODUCTION

Decision-making under partial information is a critical challenge faced by biological organisms (in particular, relating to taxis or cognition [1–4]), but also extends to broader fields such as Bayesian optimization, robotics, and machine learning [5–12]. To perform effectively, the decision-making agent must account for the physical structure of both the environment and decision problem to formulate its strategy, which often involves incorporating complex priors or tailored design elements [13–15]. To this end, general physics principles can serve as key tools, offering observables that inherently adapt to the underlying structure of the environment [16–19]. These principles are also designed to adapt to new information, even conflicting, or when the agent is limited to partial or noisy feedback from a fluctuating environment [5]. They have been applied across various domains, including modeling decision-making in the brain using the free-energy principle [20], optimizing the accuracy-complexity trade-off through the information bottleneck principle [21], and facilitating navigation in fluctuating environments via information-maximization strategies [5].

In the information-maximization framework, the agent seeks to maximize information about one or more relevant observables. The agent’s choice is thus driven by the option that maximizes the expected information gain. Due to its versatility in accommodating various environmental structures,

this principle has proven its broad applicability across both applied and theoretical areas [22,23]. Notable applications include robotics, where information sharing enhances collective decision-making [13], and taxis with the search for olfactory sources in turbulent flows [24,25]. Maximum entropy principles also underpin approaches in Bayesian optimization and reinforcement learning [10,11,14].

Multiarmed bandit (MAB) problems constitute a general framework for addressing decision-making in fluctuating and partially known environments [1,26], and have garnered significant attention in a wide range of applications, spanning from neuroscience and healthcare to recommender systems and finance [3,26–34]. In the multiarmed bandit problem, an agent is presented with a set of possible actions, or “arms,” each associated with a probabilistic reward (akin to a multiarmed slot machine game). The agent must optimize its actions to achieve a specified goal, which generally involves maximizing cumulative rewards over a fixed or infinite horizon [Fig. 1(a)]. However, it may occasionally involve identifying a subset of optimal arms or exploring its available options [Fig. 1(b)]. Hence, classic bandit settings address the exploration-exploitation dilemma and provide strategies that may achieve asymptotic optimality [35–37].

MAB problems have rapidly expanded to cover many more complex decision scenarios, and how to extend the MAB formalism and treat these generalizations is under active research. A first main class of problems focuses on pure exploration, where the agent’s goal is solely to identify a single optimal action or a subset of promising actions, without incurring penalties for sampling suboptimal ones [38–42] [Fig. 1(b)]. In essence, pure exploration aims to optimize a test phase to identify reliable candidates using a minimal amount of resources while suffering no penalty for exploration. Following such a test phase, the exploitation phase may ensue, but it is not optimized in this context. Such challenges arise in lower-risk scenarios like cosmetic drug trials or numerical

*Contact author: alex.barbier-chebbah@pasteur.fr

†Contact author: jbmasson@pasteur.fr

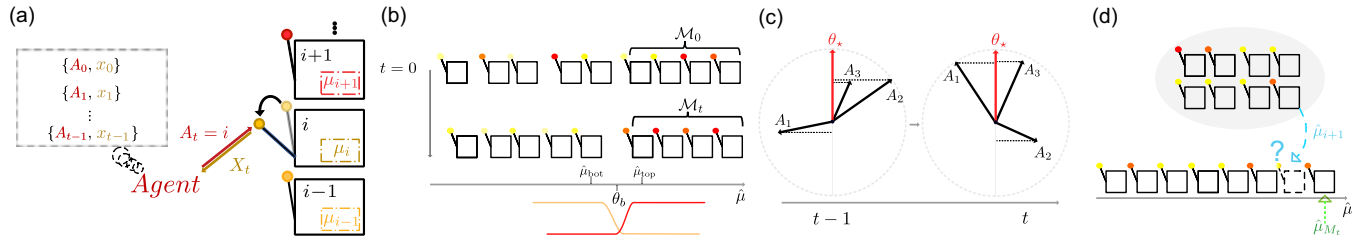


FIG. 1. Illustrations of multiarmed bandit settings addressed in this work. (a) General setting of multiarmed bandit problems. At each time step t , the agent chooses an action $i = A_t$ that returns a reward x_t drawn from a distribution of unknown mean μ_i . The arm is denoted as A_t when it is a vector and as a_t when it is an index. By accumulating rewards, the agent also gathers information about the average rewards of the arms and optimizes its decision-making policy. The specific goals of the agent will vary depending on the bandit problem. (b) General principle of the Explore- m problem. Here, the goal is to identify the m arms with the highest mean rewards (red arms). To this end, our approach introduces a well-designed separator θ_b between the current m highest arms (forming the set $\mathcal{M}_t$) and the remaining arms at each time. Our algorithm chooses the arm to maximize the information gain of the separator's effectiveness in distinguishing both subsets (cumulative distribution pictured in yellow and red below). (c) Illustration of the linear bandit setting. Here, arms are d -dimensional vectors resampled at each time step. The scalar between the arm vector and a constant but unknown vector gives the mean reward value θ_* . Because of this geometric dependency on the reward, pulling an arm also provides information on the expected results of the others. (d) Illustration of the may-arm settings. Here, the agent has a finite duration, known in advance to maximize its gains. Because there is no time to explore all the arms thoroughly, the agent needs to focus on a small subset of arms to isolate a sufficiently promising solution before time runs out. Our approach balances two possible actions: exploring a new arm (represented in blue) or greedily exploiting the current empirical best solution (represented in green) based on the empirical mean μ_{M_t} . We model the upcoming losses as a function to guide the decision-making process.

resource allocation. A second area is termed contextual bandits, which have found application, for example, in personalized recommendations and pandemic control [43,44]. The most notable example is linear bandit problems [26,45–48], where the actions' outcomes are correlated due to an underlying structure between arms, allowing one to gain information about other actions than the one performed [Fig. 1(c)]. Third, many-armed problems address scenarios where the number of available actions is large compared to the time horizon, requiring the agent to concentrate on sufficiently promising solutions instead of optimal ones [Fig. 1(d)]. This challenge of finite resources resonates with applications in pharmaceutical and medical trials [49,50]. Such settings offer various extensions and readily apply to cases where the number of arms is infinite [51–54]. Additional extensions of MAB problems explore dynamic rewards, adversarial settings, and multiagent interactions, focusing on competition, congestion, and over-exploitation, respectively, but are not investigated further here [26,42,45,55–57].

Although a wide variety of bandit games and associated strategies have been studied in recent years, strategies developed to address them primarily rely on a few core principles, most notably confidence bounds on the individual arms' mean rewards, which frequently drive the decision-making process [37,58–60]. Although there are exceptions to this rule and several alternative approaches have been investigated [36,48,55,61–64], there is still a crucial need to develop original approaches, especially when addressing more complex bandit problems.

Recently, the information-maximization principle has been shown to achieve strong results in classical bandit settings, displaying state-of-the-art empirical performance both at short and long time horizons and when handling multiple arms [62,65]. Information-maximization algorithms (and information-based algorithms, more generally) aim to minimize the expected entropy (a functional) of a key variable within the system.

The first difficulty in applying information-based algorithms is that they rely on integrating a function that often has no known analytical solution. Their application thus relies on numerical integration, which both hinders theoretical analysis of their statistical performance and increases their computational costs [65]. Recently, we introduced a different class of approximate information-maximization (AIM) algorithms [65,66] that rely on tractable approximations of the entropy. The analytical tractability of the AIM algorithms significantly increases their computational efficiency, allowing their application to a wider range of bandit problems [65,66]. Furthermore, it enabled us to rigorously prove their asymptotic optimality in the classic setting with Gaussian and sub-Gaussian rewards [66], confirming the heuristic observation that information maximization can drive optimal learning and optimization.

A second challenge with information-based approaches remains choosing the observable and functional to appropriately capture the underlying structure of the bandit problem and ensure an efficient policy. The main challenge of information maximization is to avoid overexploration [62], requiring careful design of the observables or features on which entropy minimization is performed as well as of the approximations performed to make the algorithms tractable [65].

In this work, we show how the approximate information-maximization principle can be applied to the three distinct bandit topics discussed above. We, namely, consider the pure exploration problem of explore- m selection, contextual bandits exemplified by linear bandits, and many-armed bandits. The three problems each involve distinct levels of exploration, and we discuss how relevant information is accessed within each context. This allows us to extend the information-maximization principle to a broader range of structures not specifically meant for pure exploration. We provide numerical evaluations of the performance of the approximate information-based approaches and compare it to well-established algorithms in each case. A key challenge

left for future work is to establish theoretical bounds for the algorithms investigated below.

A. Structure of this paper

In Sec. II, we briefly introduce the classic MAB problem and review information-maximization approaches to deal with this problem. We detail in the following sections our approach to the three distinct bandit settings mentioned above, with Sec. III focusing on the Explore- m problem, Sec. IV on linear bandits, and Sec. V on the many-armed bandit problem. For each case, we follow the same three steps: (1) we introduce the bandit game, defining the key elements that are required of an efficient strategy, and provide an overview of existing strategies to tackle the problem; (2) we detail our design principles for an information-based approach, followed by the development of our approximation scheme that yields a tractable expression and enables an efficient implementation; and (3) finally, we investigate the performance of the resulting algorithm, its specific features, and we discuss potential extensions.

Several Appendixes provide detailed derivations of our proposed algorithms (Appendixes A–C) as well as numerical experiments for benchmarking and comparing them to existing baselines (Appendix D).

II. MULTIARMED BANDIT PROBLEM

In the classic multiarmed stochastic bandit game, an agent iteratively selects an action, or *arm*, among a predefined set of K possibilities. In each round, indexed by t , the agent selects an arm $i_t \in \{1, \dots, K\}$ based on the rewards of the previously pulled arms [Fig. 1(a)]. The chosen arm, i_t , then returns a stochastic reward x_t , drawn independently of the previous rounds, according to a distribution v_i of unknown, but fixed, mean μ_i . We denote by t_i the number of times the i th arm has been drawn and by $\mathcal{H}_t = \{(i_1, x_1), (i_2, x_2), \dots, (i_t, x_t)\}$ the history of the game (the drawn arms and their rewards) after t rounds. For simplicity, we omit the dependence on t when this is clear from the context.

The agent's goal is to maximize its expected cumulative reward over T rounds. This is equivalent to minimizing the expected value of the *regret*, which is defined as [26]

$$R_T = T\mu_* - \sum_{t=1}^T \mathbb{E}[\mu_{i_t}], \quad (1)$$

where $\mu_* = \max_{i \in [K]} \mu_i$. This measures the gap between the cumulative reward from always choosing the best arm and the agent's expected reward over T rounds. The expectation reflects that the agent's policy, informed by past rewards, is affected by their randomness in a single game.

The expected regret is asymptotically lower bounded by the Lai-Robbins bound [35],

$$\lim_{T \rightarrow \infty} \frac{R_T}{\ln(T)} \geq \sum_{k: \mu_k < \mu_*} \frac{\mu_* - \mu_k}{D_{\text{KL}}(v_k \| v_{k^*})}, \quad (2)$$

where $k^* = \operatorname{argmax}_{i \in [K]} \mu_i$ and $D_{\text{KL}}(v_k \| v_{k^*})$ is the Kullback-Leibler divergence between the reward distributions of the arms k and k^* . The right-hand side of Eq. (2) is independent

of T , so the equation shows that the expected regret of any decision strategy grows at least logarithmically with T for long times.

A bandit algorithm that asymptotically attains the Lai-Robbins bound [Eq. (2)] for given set of reward distributions $[v_i]$ is termed optimal. Many algorithms have been devised to tackle the classic MAB problem, notably the families of ϵ -greedy strategies [67], upper confidence bound (UCB) algorithms [58], and Thompson sampling [61], which have all been demonstrated to be able to attain the Lai-Robbins bound in a range of settings. However, the complexity of their implementation and their efficiency at finite times may vary significantly.

Information maximization [62], inspired by principles from statistical physics and information theory, has emerged as another viable strategy for tackling the MAB problem [65,66]. The strategy relies on optimizing a single functional in each round of the bandit game. In particular, for the classic MAB problem, information maximization aims to minimize the entropy,

$$S = - \int p_{\max}(\theta) \ln p_{\max}(\theta) d\theta, \quad (3)$$

of the posterior distribution $p_{\max}$ of the value of the maximal reward, μ_* . In practice, the strategy chooses the arm that minimizes the expected increment of the entropy,

$$\operatorname{argmin}_{i \in [K]} \mathbb{E}[S_{t+1}(i) - S_t \mid i_{t+1} = i], \quad (4)$$

where S_t is the entropy in round t and $S_{t+1}(i)$ the entropy in the next round if i were drawn. Numerical experiments have shown information maximization to be highly efficient at short and intermediary times [62,65].

Evaluating the entropy S [Eq. (3)] analytically is in general intractable, which makes exact information maximization computationally demanding and difficult to analyze analytically. To overcome this challenge, we developed an AIM strategy [65], which was recently shown to be asymptotically optimal for sub-Gaussian reward distributions [66]. We explore below how to adapt the approximate information-maximization principle to more general bandit problems.

III. EXPLORE- m BANDITS

A. Setting and core insights

We first consider a well-known multiple-arm identification problem, commonly referred to as the Explore- m problem under the probably approximately correct (PAC) identification formulation [41,68–71]. In these problems, we aim to identify a subset of m promising arms (i.e., individual actions or choices) among n possible options [Fig. 1(b)]. Each arm is associated with a reward distribution of unknown mean parameter μ_i (supported on $[0, 1]$ without loss of generality). At each time step t , the agent selects an arm i_t and obtains an associated stochastic reward x_t sampled from the i th arm's reward distribution. By sampling the environment, the agent sequentially accumulates information on the parameters of the sampled arms' means μ_i . For definiteness, we sort the arms according to their (unknown) means $\mu_n \leq \mu_{n-1} < \dots < \mu_m < \dots < \mu_1$.

Explore- m aims to identify the subset of m arms among the highest means up to a predefined constant ϵ , which is thus termed ϵ -optimal. More precisely, an arm i is defined to be ϵ -optimal, if and only if $\mu_i \geq \mu_m - \epsilon$ for some $\epsilon \in (0, 1)$ where μ_m is the m th highest mean. This criterion prevents the algorithm from spending excessive time distinguishing between arms when μ_m and μ_{m+1} are infinitesimally close, ensuring it focuses on more meaningful scenarios.

Since no additional cost is incurred by sampling suboptimal arms, the algorithm's efficiency is evaluated based solely on its ability to identify such a subset both quickly and with high confidence. Two standard evaluation criteria are commonly considered [40]. In the fixed-budget setting, the algorithm has a finite amount of time to identify the subset, and its performance is measured through the probability of returning an ϵ -optimal subset. In contrast, the PAC formulation requires the algorithm to stop when the probability of returning an ϵ -optimal subset exceeds $1 - \delta$, with $\delta \in (0, 1)$, with the goal of minimizing the stopping time [41]. We focus on the PAC formulation in the following.

Algorithms for pure exploration problems can broadly be divided into two categories. In *sample and reject strategies*, a subset of remaining arms are sampled at each round. Next, a rejection criterion discards from the remaining subset sufficiently drawn arms. This process is repeated sequentially until a stopping criterion is met [54,72]. For *adaptive sampling strategies*, at each time step, one or two arms with the most critical confidence regarding the subset identification are selected and pulled, with the arms' indices changing over time [40,41,73]. Most methods rely on upper and lower bounds on confidence intervals on the arms' means, rooted in the general class of UCB algorithms [39,72–74]. By comparing the upper confidence bounds of likely suboptimal arms with the lower confidence bounds of the promising subset, these methods can derive an efficient decision-making procedure. However, the stopping criteria are highly dependent on the confidence interval design, which may become less effective when facing various reward distributions or more complex scenarios. For a deeper discussion of the sampling complexity of such strategies under different reward distributions, we refer to Refs. [68,70].

Since no explicit cost is incurred by selecting arms, we would expect a strategy focused solely on acquiring information about the best subset to be efficient. To this end, we develop a fully information-based strategy focused on the PAC setting, termed probably approximately correct approximate information maximization (PacAIM), which ensures feasibility and tractability through successive derivations, approximations, and simplifications. While we focus here on the PAC setting, we discuss its extension to other pure exploration settings at the end of the section.

B. Design principle

We first introduce additional notation specific to the Explore- m problem. In each round t , each arm i is characterized by the number of times it has been drawn t_i and its empirical mean $\hat{\mu}_i = \sum_{t:i=i} x_t / t_i$. We denote by $\mathcal{H}_t = \{(i_1, x_1), (i_2, x_2), \dots, (i_t, x_t)\}$ the complete history of chosen arms and observed rewards. The subset of arms with the

current m highest empirical means is denoted by $\mathcal{M}_t$, with associated sets of true (unknown) mean values, $\{\mu_{\mathcal{M}_t}\}$, and empirical means, $\{\hat{\mu}_{\mathcal{M}_t}\}$. The equivalent notations for the remaining arms are $\mathcal{D}_t = \{n\} \setminus \mathcal{M}_t$, $\{\mu_{\mathcal{D}_t}\}$, and $\{\hat{\mu}_{\mathcal{D}_t}\}$, respectively. We also denote the worst selected mean by $\hat{\mu}_{\text{top}} = \min(\{\hat{\mu}_{\mathcal{M}_t}\})$ and the best unselected by $\hat{\mu}_{\text{bot}} = \max(\{\hat{\mu}_{\mathcal{D}_t}\})$.

A promising candidate distribution for information maximization is the probability of $\mathcal{M}_t$ being ϵ -optimal, i.e., $\mathbb{P}_{\mathcal{H}_t}(\min(\{\mu_{\mathcal{M}_t}\}) > \max(\{\mu_{\mathcal{D}_t}\}) - \epsilon)$, where $\mathbb{P}_{\mathcal{H}_t}$ denotes the conditional probability given the history of the previous draws, $\mathcal{H}_t$ [75]. Unfortunately, this posterior distribution is analytically intractable in practice. Additionally, since the arms are not identically distributed, applying asymptotic extreme value statistics is also challenging without approximations.

Therefore, we adjust the decision process as follows. We introduce θ_b as a separator value between the observed sets $\{\hat{\mu}_{\mathcal{M}_t}\}$ and $\{\hat{\mu}_{\mathcal{D}_t}\}$ and compare all the arms' mean likelihoods relative to θ_b . In other words, we now focus on comparing the arms' posterior probabilities of their means being above or below θ_b , given by

$$\begin{aligned} P_{\text{top}} &= \mathbb{P}_{\mathcal{H}_t}(\min(\{\mu_{\mathcal{M}_t}\}) > \theta_b - \frac{\epsilon}{2}) \\ &= \prod_{i \in \mathcal{M}_t} \mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2}) \end{aligned} \quad (5)$$

and

$$\begin{aligned} P_{\text{bot}} &= \mathbb{P}_{\mathcal{H}_t}(\max(\{\mu_{\mathcal{D}_t}\}) < \theta_b + \frac{\epsilon}{2}) \\ &= \prod_{i \in \mathcal{D}_t} \mathbb{P}_{\mathcal{H}_t}(\mu_i \leq \theta_b + \frac{\epsilon}{2}), \end{aligned} \quad (6)$$

respectively. Notably, if both events occur, $\mathcal{M}_t$ is ϵ -optimal, making Eqs. (5) and (6) an effective proxy to derive an information-maximization strategy (see Appendix A for the link between the entropy and posterior optimization). This leads us to a policy that selects the arm maximizing the expected increase in $P_{\text{bot}}P_{\text{top}}$. Since the arms are independent, we can derive a straightforward expression for the expected increase if arm i were selected (see Appendix A for details). The expression differs for arms in $\mathcal{M}_t$ and $\mathcal{D}_t$ and is given by

$$\Delta_{i \in \mathcal{M}_t} P_{\text{top}} P_{\text{bot}} = \frac{\Delta \mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2})}{\mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2})} P_{\text{top}} P_{\text{bot}}, \quad (7)$$

for $i \in \mathcal{M}_t$, and

$$\frac{\Delta_{i \in \mathcal{D}_t} P_{\text{bot}} P_{\text{top}}}{P_{\text{top}} P_{\text{bot}}} = \frac{\Delta \mathbb{P}_{\mathcal{H}_t}(\mu_i \leq \theta_b + \frac{\epsilon}{2})}{\mathbb{P}_{\mathcal{H}_t}(\mu_i \leq \theta_b + \frac{\epsilon}{2})}, \quad (8)$$

for $i \in \mathcal{D}_t$, and where Δ denotes the expected discrete increment along arm i . Exact expressions can be derived for both choices (see Appendix A 2). The selected arm is then given by

$$a_t = \operatorname{argmax}_{i \in \mathcal{M}_t \cup \mathcal{D}_t} (\Delta_i P_{\text{bot}} P_{\text{top}}). \quad (9)$$

In other words, we draw the arm with the largest relative increment at θ_b .

This approach offers two main advantages. First, the information gain is significantly simplified as it relies on a single-point evaluation. Second, it directly provides a well-defined stopping criterion for the PAC setting. As noted above,

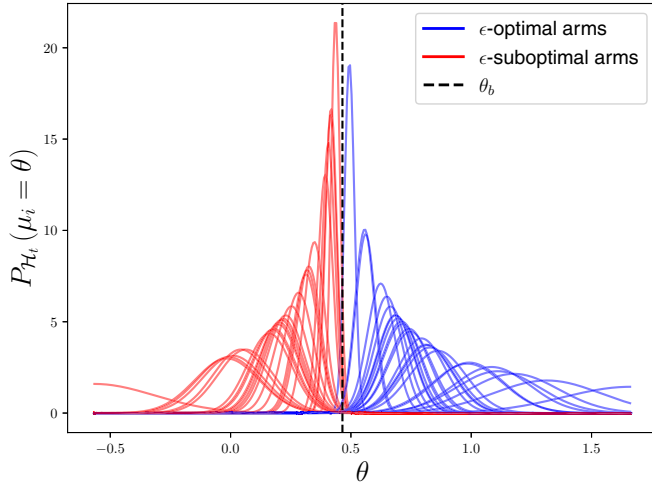


FIG. 2. Posterior distributions of the arms' mean at the stopping time. The individual posterior distribution for each arm mean is shown at the stopping time with our PacAIM for $\epsilon = 0.02$, $\delta = 0.1$, $m = 25$, and $n = 50$. The ϵ -optimal arms, among which m should be returned, are shown in blue, while the remaining arms are displayed in red. The θ_b value, deduced from Eq. (A20), is indicated by a vertical line. Distributions closer to θ_b are narrower, indicating that the algorithm more frequently selects arms near the threshold. See Appendix E 1 for how the number of draws of each arm depends on its estimated mean.

the joint probabilities [Eqs. (5) and (6)] lower bound the probability of returning an ϵ -optimal subset. As both P_{top} and P_{bot} are easy to evaluate, they can be directly translated into a stopping criterion. That is, the algorithm stops when

$$P_{\text{bot}}P_{\text{top}} > 1 - \delta \quad (10)$$

is satisfied for the first time. Hence, the decision-making process is designed to optimize the stopping time.

The final step needed for an efficient strategy consists in finding a good θ_b value. We seek an optimal separator, i.e., the θ_b value that enables fulfilling Eq. (10) in minimal time. Introducing the total time needed to be spent on sampling each subset in order to verify Eq. (10) (regarding the current arms' mean state), T_{top} and T_{bot} , respectively, θ_b can be expressed as

$$\theta_b = \underset{\theta \in [\hat{\mu}_{\text{bot}}, \hat{\mu}_{\text{top}}]}{\text{argmin}} (T_{\text{bot}}(\theta_b) + T_{\text{top}}(\theta_b)) | P_{\text{bot}}P_{\text{top}} > 1 - \delta. \quad (11)$$

Unfortunately, there is no exact and tractable solution for Eq. (11). Hence, we seek tractable approximations of θ_b , $T_{\text{bot}}(\theta_b)$, and $T_{\text{top}}(\theta_b)$. To this end, we isolated from the posterior distributions of the maximum and the minimum, respectively, of $\{\mu_{\mathcal{M}_t}\}$ and $\{\mu_{\mathcal{D}_t}\}$, the following heuristic closed-form approximation for each set (see Figs. 2 and 9). We used an extreme value analysis to approximate P_{bot} and P_{top} by Gumbel laws parametrized by $\hat{\mu}_{\text{top}}$, $\hat{\mu}_{\text{bot}}$, $T_{\text{bot}}(\theta_b)$, and $T_{\text{top}}(\theta_b)$:

$$\theta_b = \frac{\epsilon}{2} + \hat{\mu}_{\text{top}} - \sqrt{\frac{m}{T_{\text{top}}(\theta_b)}} \left[\frac{2 \ln(m) - \ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(m)}} \right] \quad (12)$$

ALGORITHM 1. Algorithm for Explore- m Gaussian bandits.

Data: $n, m, \epsilon, \delta, T > n$ (optional)

Draw all arms once, observe rewards and update statistics for all arms, $t \leftarrow n$

continue $\leftarrow$ **True**

while *continue* **do**

$\mathcal{M}_t = \underset{I \subseteq \{1, \dots, n\}, |I|=m}{\text{argmax}} (\hat{\mu}_i)$ and $\hat{\mu}_{\text{top}} = \max\{\hat{\mu}_{\mathcal{M}_t}\}$

$\mathcal{D}_t = \{n\} \setminus \mathcal{M}_t$ and $\hat{\mu}_{\text{bot}} = \max\{\hat{\mu}_{\mathcal{D}_t}\}$

Evaluate P_{top} and P_{bot} using Eqs. (5) and (6)

if Eq. (10) **or** $t = T$ **then** *continue* $\leftarrow$ **False**
 // Stopping criterion

else Evaluate θ_b using Eq. (11)

Select a_{t+1} with Eq. (9) // Arm selection

Pull a_{t+1} and observe $X_{t+1}(a_{t+1})$

return $\mathcal{M}_t$ // ϵ -optimal set

and

$$\theta_b = -\frac{\epsilon}{2} + \hat{\mu}_{\text{bot}} + \sqrt{\frac{n-m}{T_{\text{bot}}(\theta_b)}} \times \left[\frac{2 \ln(n-m) - \ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(n-m)}} \right]. \quad (13)$$

By inserting Eqs. (12) and (13) into Eq. (11), we obtain a closed system from which we derive a tractable expression of θ_b given in Eq. (A20). We refer to Appendix A for a complete and detailed review of all the approximations and derivations.

Even though this approximation of θ_b may not be asymptotically optimal for all mean priors, it empirically provides an efficient stopping criterion as compared to traditional upper confidence bound approaches.

Pseudocode for the resulting algorithm referred to as PacAIM is presented in Algorithm 1. Figure 2 shows estimated posterior distributions of each arm's mean and of the separator value θ_b after a single run of the algorithm, illustrating how the arms closer to the boundary between the optimal and suboptimal sets are drawn much more, leading to more concentrated posteriors for these arms.

C. Results

To evaluate the performance of PacAIM, we compare in Fig. 3 its stopping performance with the LUCB1 algorithm [41], a well-established algorithm in the PAC setting. The reward distributions are Gaussian with unknown mean values in $[0, 1]$ and unit variance.

In the first experiment, the ϵ -optimal subset is formed by mean rewards drawn uniformly from $[0.9, 1]$, while the remaining mean rewards are drawn from $[0.0, 0.8]$. To assess the algorithm's performance on more complex tasks, we also evaluate a scenario where all ϵ -optimal rewards are set to 0.8 and the remaining mean rewards are set to 0.7. The stopping time is measured for different numbers of arms while maintaining the ratio $m = n/5$ (see Appendix D for details on numerical simulations).

In both cases, PacAIM outperforms LUCB1, demonstrating the ability of our approximation to accurately capture the

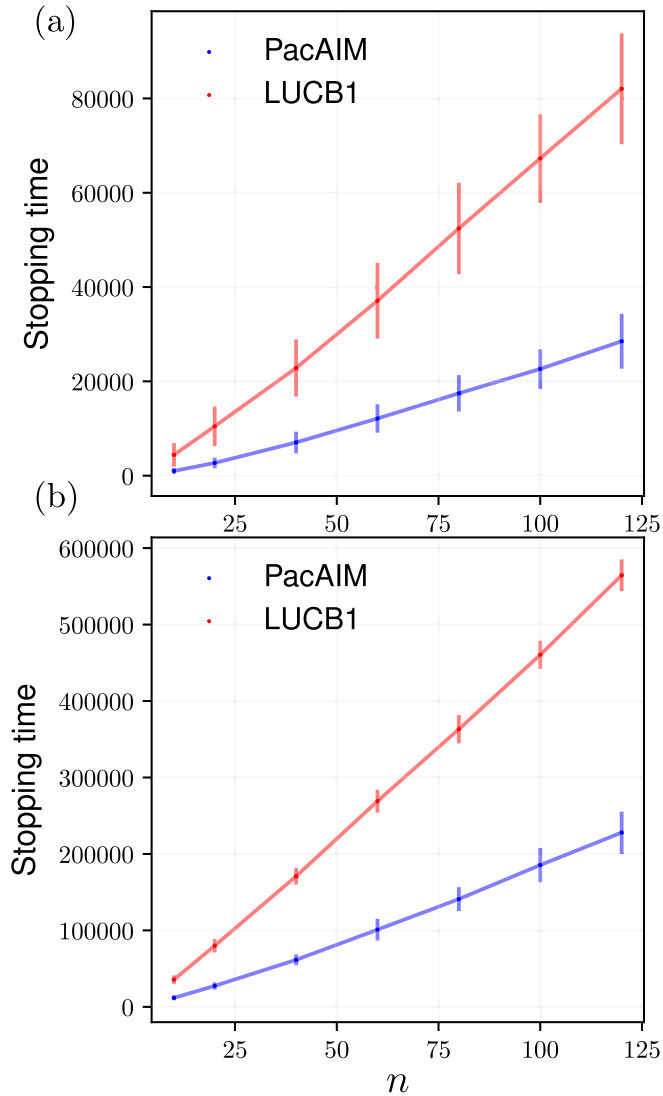


FIG. 3. Stopping time of Explore- m algorithms for Gaussian rewards. The stopping time, i.e., the time when the stopping criteria for each algorithm are met, is measured for different numbers of arms n . In blue our algorithm, and in red LUCB1 given in Ref. [41]. The ϵ -optimal subset size verifies $m = n/5$. In panel (a), the mean reward values are sampled from two disjoint ensembles $[0, 0.8]$ and $[0.9, 1]$ for the ϵ -optimal subset, with confidence parameters $\epsilon = 0.05$ and $\delta = 0.1$. (b) Suboptimal mean reward values are all set to 0.7, while ϵ -optimal to 0.8, with confidence parameters $\epsilon = 0.05$ and $\delta = 0.02$. The time is averaged over 100 runs with standard deviation of the stopping time indicated (see Appendix E 1 for a comparison of the success probabilities of the PacAIM and LUCB1 algorithms). Details on numerical simulations and on the LUCB1 algorithm are found in Appendix D.

significant contributions of all arms near the θ_b value. Notably, we emphasize that using a decision procedure directly tied to the stopping criteria enables the algorithm to sample all relevant arms efficiently, minimizing the stopping time. Additionally, we verified that our solution still achieves a success probability exceeding the tolerance threshold $1 - \delta$, thereby validating both our solution and the obtained stopping time (see Fig. 8, Appendix D).

To conclude on the Explore- m setting, we highlight some additional observations. First, θ_b depends on the current state of the game and evolves over time to adapt to the reward distributions. Next, the algorithm adjusts by sampling the arms that provide the most information relative to this updated separator value.

It is worth noting that our expression is designed for a subset consisting of multiple arms. For single-arm identification, one may replace Eq. (12) by directly expressing θ_b in terms of T_{top} using $\mathbb{P}_{\mathcal{H}_t}(\hat{\mu}_{\text{top}} \geq \theta_b - \frac{\epsilon}{2}) = 1 - \delta/2$. Furthermore, the expressions in Eqs. (12) and (13) are derived for Gaussian rewards, but equivalent formulations can be obtained for various distribution families [76].

As a result, our algorithm should be adaptable to more complex reward distributions, including ones with power-law tails, which lead to extreme value statistics scaling functions, such as Fréchet distributions [76]. Altogether, these first results highlight the potential for extending information-based strategies to tackle more general pure exploration problems [38,77,78]. The design of approximations and the choice of observables should be addressed in a problem-specific manner. However, our work can be easily adapted to threshold identification or gap identification problems [79,80], where the focus is on determining the gap between subsets. Additionally, our approach may be extended to identify all arms above a certain threshold or to target a specific fraction of arms, such as identifying all ϵ -good arms or for detecting outliers [81–83]. In all these examples, we expect information-based strategies to prove efficient in capturing the structure and correlations within the game [38,77].

IV. LINEAR BANDITS

A. Setting and core insights

Contextual bandits with linear payoffs are another paradigmatic extension of the classic MAB problem, underpinning key applications such as internet advertisement selection and personalized recommendations [46,84,85]. This problem extends the classic bandit framework by introducing correlations between arms and allowing the set of arms to change over time [26,46,48,86] [Fig. 1(c)]. We here focus on the stochastic linear problem. At each round t , the agent faces the decision set $\mathcal{A}_t \subset \mathbb{R}^d$, which may vary over time. From this set, it selects an action A_t and observes a corresponding reward,

$$X_t = \langle \theta_*, A_t \rangle + \xi_t, \quad (14)$$

where θ_* defines the optimal direction and ξ_t is modeled as Gaussian white noise in this study. The mean reward value is linear with respect to A_t , inducing a correlation between arms that tends to favor actions aligned with θ_* . Crucially, the reward at round t is determined by the action-context pair, rather than by any intrinsic identity of an arm. Arm labels are therefore arbitrary; the decision set $\mathcal{A}_t$ serves as context by specifying the available actions. This context dependence is precisely why such settings are designated as contextual bandit problems, within which the linear bandit is a canonical example.

Since the decision set may evolve over time, it is more convenient to evaluate the losses based on the best option available at any given time, leading to the pseudoregret

[26,87]:

$$\hat{R}_n = \sum_{t=0}^n \max_{a \in \mathcal{A}_t} (\theta_*, a - A_t), \quad (15)$$

which is a random variable with expectation $R = \mathbb{E}[\hat{R}_n]$.

Heuristically, as the agent continuously adapts to the current decision set, $\mathcal{A}_t$, it aims to gather information about θ_* —the key element shared across all decisions—while avoiding actions that are orthogonal or opposite to it. In the d -dimensional space, certain directions may be underexplored, which necessitates more exploration at the condition of an appropriate trade-off. Hence, the composition of $\mathcal{A}_t$ plays a crucial role in the setting. If $\mathcal{A}_t$ forms a constant orthogonal basis, it simplifies to the classical bandit setting. The decision set may be finite or infinite, generated by a stochastic process, or part of an adversarial game adapting to previous choices, with $\mathcal{A}_t$ potentially evolving rather than being drawn from a predefined set of decision sets [26]. This inherent diversity makes it difficult for linear bandit algorithms to be optimal across all settings [47,88].

As in most bandit problems, the UCB algorithm has emerged as a useful approach. It relies on a tuned confidence index associated with each arm to decide which arm to play. In the linear bandit case, the core algorithm is referred to as LinUCB, which achieves a logarithmic scaling of the regret when actions sufficiently close to θ_* are available [26,43,85,89–92]. Otherwise, in the general case, the optimal regret is bounded by $O(d \sqrt{T})$. Beyond the UCB class, Thompson sampling, which relies on sampling mean rewards from a posterior distribution and choosing the arm with probability proportional to this posterior, has also been shown to achieve near-optimal regret bounds for stochastic contextual bandits with linear payoff functions [36,48,93,94]. Optimal allocation matching strategies, on the other hand, rely on an estimate of θ_* and solve an optimization problem based on this estimate or perform exploration to determine the next action [47,64,88,91]. This class of strategies has been shown to achieve optimality when faced with a fixed action set [88,91]. Finally, the information-directed sampling (IDS) principle leverages a measure of information gain from the optimal action, explicitly balanced by expected losses from exploration. By consistently using an information measure that captures the specific problem structure, IDS has been shown to achieve optimal regret scaling for linear bandits and other structured settings [9,95–97]. As first found in Ref. [47], a key insight is that a consistent strategy must allocate exploration time so that the gaps between all suboptimal actions and θ_* are identified with high confidence. Otherwise, a purely greedy strategy solely based on θ_* confidence estimation will fall short. This highlights the complexity arising from the variation of $\mathcal{A}_t$.

We here develop a principle that relies almost exclusively on information acquisition to drive its decision policy. Our approach is rooted in the AIM strategy [65,66], which involves selecting the arm that maximizes the information gained from a judiciously chosen observable of the game. As reviewed in Sec. II, this gain is measured through the associated entropy, with its increments approximated to identify the arm that will result in the largest entropy reduction. In the

classic MAB problem, the entropy of the posterior distribution of the maximum mean reward value results in an optimal exploration-exploitation trade-off. However, it cannot be directly mapped to the linear case since, by symmetry, A_{M_t} and $-A_{M_t}$ carry the same information. Thus, suboptimal actions can provide extensive information on θ_* value but incur higher losses. Hence, solely focusing on acquiring more information on θ_* results in overexploration. We now present an adapted version of AIM for the linear bandits problem, denoted by LinAIM for short, to address this challenge.

B. Algorithm design

Similar to Explore- m bandits, we first define relevant notation. Let $B_t = \lambda I_d + \sum_{n=1}^t A_n A_n'$ be the inverse of the covariance matrix of the posterior distribution of θ_* according to a symmetric Gaussian prior parametrized by λ . We denote by $\hat{\theta}_t = B_t^{-1} \sum_{n=1}^t X_n A_n$ its associated posterior mean (see Appendix B for details). Here, $'$ and $^{-1}$ represent the transpose and inverse operators, respectively. The best available empirical arm in $\mathcal{A}_t$ is $A_{M_t} = \max_{a \in \mathcal{A}_t} \langle \hat{\theta}_t, a \rangle$. We finally define a norm through $\|x\|_G = x' G x$ for $x \in \mathbb{R}^d$ and with $G \in \mathbb{R}^{d \times d}$ positive definite.

To adapt AIM, we opt to adjust the information gain depending on the selected arm. Arms will be selected to maximize the information regarding the value of θ_* , weighted by the likelihood of the arm being suboptimal. In other words, the arms are evaluated not only by the information they provide but also regarding their relative gap with A_{M_t} , quantified by the likelihood that the best empirical choice A_{M_t} is better than A_t , i.e., the probability that $\langle \theta_*, A_{M_t} - A_t \rangle > 0$. Finally, A_{M_t} is assigned to maximize the information regarding the value of θ_* . This procedure offers two main benefits: (1) it prevents overexploration by adjusting the information gains from suboptimal arms. (2) It also involves a measure of the gaps between mean rewards, addressing the criteria set out by Ref. [47], which are also implicitly included in the information regarding θ_* .

We now provide the analytical expression regarding the entropy increments relative to each choice (see Appendix B for details).

We first consider the information gain tied to A_{M_t} . The posterior distribution of θ_* is given by

$$p(\theta_* = \theta | \mathcal{H}_t) = \frac{\exp(-\frac{1}{2}(\theta - \hat{\theta}_t)' B_t (\theta - \hat{\theta}_t))}{\sqrt{(2\pi)^d \det B_t^{-1}}}, \quad (16)$$

where $\mathcal{H}_t$ represents the history of chosen arms and observed rewards. The associated entropy can be expressed exactly as

$$S = \frac{1}{2} \ln(\det B_t^{-1}) + \frac{d}{2} [1 + \ln(2\pi)]. \quad (17)$$

The corresponding entropy increment, which quantifies the gain of information about θ_* , is equal to

$$\Delta S_{A_{M_t}} = -\frac{1}{2} \ln(1 + A_{M_t}' B_t^{-1} A_{M_t}). \quad (18)$$

We now consider the remaining arms, focusing on a given suboptimal arm A_t . The probability of observing

ALGORITHM 2. Algorithm for d -dimensional linear Gaussian bandits.

Draw a first arm A_0 , observe reward X_0 , and update statistics: $B_1 = \lambda I_d + A_0 A_0'$ and $\hat{\theta}_1 = B_1^{-1} X_0 A_0$

for $t = 1$ **to** T **do** // Arm selection

$A_{M_t} \leftarrow \max_{a \in \mathcal{A}_t} \langle \hat{\theta}_t, a \rangle$

Evaluate $\Delta S_{A_{M_t}}$ following Eq. (18)

for $A_n \in \mathcal{A}_t \setminus \{A_{M_t}\}$ **do**

Evaluate $\Delta \tilde{S}_{A_n}$ following Eq. (22)

Select A_t as

$\operatorname{argmin} \left(A_n \in \mathcal{A}_t \setminus \{A_{M_t}\}, A_{M_t} : \Delta \tilde{S}_{A_n}, \Delta S_{A_{M_t}} \right)$

Pull A_t and observe $X_t(A_t)$

$B_{t+1} \leftarrow B_t + A_t A_t'$ and

$\hat{\theta}_{t+1} \leftarrow B_{t+1}^{-1} \sum_{n=1}^t X_n A_n$ // Update statistics

$\langle \theta_*, A_{M_t} - A_t \rangle > 0$, termed $C_{A_{M_t}, A_t}$, is given by

$$C_{A_{M_t}, A_t} = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\theta_*' [A_{M_t} - A_t]}{\sqrt{2 [A_{M_t} - A_t]' B_t^{-1} [A_{M_t} - A_t]}} \right) \right], \quad (19)$$

with associated entropy

$$S_{C, A_t} = -C_{A_{M_t}, A_t} \ln C_{A_{M_t}, A_t} - (1 - C_{A_{M_t}, A_t}) \ln (1 - C_{A_{M_t}, A_t}). \quad (20)$$

Thus far, all steps have been computed exactly, but to achieve a simpler expression, we focus on the asymptotic form of Eq. (20) when A_{M_t} is assumed to be the most probable optimal option. We denote this asymptotic form by w_{A_t} such that $S_{C, A_t} \sim w_{A_t}$ as $\frac{\delta}{\sqrt{\Sigma}} \rightarrow \infty$, which takes the form

$$w_{A_t} = \frac{\delta}{\Sigma^{\frac{3}{2}}} \frac{e^{-\frac{\delta^2}{2\Sigma}}}{2\sqrt{\pi}}, \quad (21)$$

with $\delta = \theta_*' [A_{M_t} - A_t]$ and $\Sigma = [A_{M_t} - A_t]' B_t^{-1} [A_{M_t} - A_t]$ (see Appendix B). The weighted information gain relative to A_t thus reads

$$\begin{aligned} \Delta \tilde{S}_{A_t} &= -w_{A_t} \Delta S_{A_t} \\ &= -\frac{1}{2} \frac{\delta}{\Sigma^{\frac{3}{2}}} \frac{e^{-\frac{\delta^2}{2\Sigma}}}{2\sqrt{\pi}} \ln (1 + A_t' B_t^{-1} A_t). \end{aligned} \quad (22)$$

The pseudocode for the LinAIM algorithm is presented in Algorithm 2.

C. Results

Suboptimal arms are evaluated regarding their gap with the best empirical arm available. As a result, the information gain expressed by Eq. (21) decreases exponentially with respect to this gap, making Eq. (18) dominate in most cases, thereby preventing our algorithm from overexploring. In particular, it makes Eq. (21) asymmetric with respect to A_t , thereby enabling arms of opposite signs to be told apart. This fulfills the key challenge of preventing suboptimal arms from being sampled regarding the information gain they provide

for identifying optimal solutions. It also introduces the information relative to the gaps between current arms' means. Nonetheless, suboptimal arms are still evaluated based on the information they provide on the full system and under probed directions, which stays the core of AIM.

A preliminary theoretical guarantee can be assessed by considering the classic bandit setting, where $\mathcal{A}_t$ is constant and forms an orthogonal basis. In this setting, Eqs. (18) and (21) demonstrate asymptotic behavior similar to previous information-maximization algorithms [66] that have recently been proven optimal. A second illustrative example involves an adversarial scenario where $\mathcal{A}_t$ will suddenly start to display solely actions that have not been explored previously—untested directions, i.e., with high values of $A_{M_t}' B_t^{-1} A_{M_t}$. In such cases, the entropy increment [Eq. (18)] should dominate, still accounting for exploration despite facing solely highly uncertain options.

To evaluate the performance of LinAIM, the average regret is analyzed in Fig. 4 for two complementary settings meant to cover a wide range of scenarios. In the first experiment, $\mathcal{A}_t$ is formed by ten arms resampled at each time from a normal distribution in $\mathbb{R}^{10}$. This decision set exhibits temporal variability, with a sparse number of solutions available at each discrete time step. Here, LinAIM performances match those of commonly used algorithms such as linUCB [26,90], slightly outperforming them. It showcases the algorithm's efficiency when confronted with a broad range of arms. To evaluate LinAIM's ability to leverage available information in suboptimal solutions, we consider a toy setting defined in Ref. [64] originally from Ref. [47]. The agent faces two fixed sets $\mathcal{A}_t$ formed by three arms in $\mathbb{R}^3$, each set having an equal probability of being presented to the agent. The first set is $\{[1, 0, 0], [0, 1, 0], [1 - \xi, 2\xi, 0]\}$, the second set $\{[0, 0.6, 0.8], [0, 0, 1], [0, \xi/10, 1 - \xi]\}$, and finally $\theta_* = [1, 0, 1]$ with ξ small. The key insight is that optimistic strategies rarely select $[0, 1, 0]$ and $[0, 0.6, 0.8]$, respectively, in the first and second sets due to the substantial gap. They thereby incur significant regret in learning which of the remaining arms is optimal. Conversely, an asymptotically optimal strategy strategically allocates more selections to the suboptimal arms, facilitating information acquisition to accurately identify θ_* , hence avoiding a regret that scales with ξ . Here again, our AIM-based algorithm shows strong performance similar to that of state-of-the-art algorithms. Such performances are quite convincing, and our algorithm demonstrates performance comparable to recent algorithms such as information-directed sampling [95,97], SOLID [64], OSSB [88], or OAM [91]. Although our algorithm shares similarities with information-directed sampling by adjusting the information gain, our design and analytic expressions are simpler than the methods cited above. It notably reduces the implementation costs and does not require working with finite arm sets. Nevertheless, the linear bandit setting is highly complex, particularly in the adversarial regime where the decision set can be specifically designed to expose the algorithm's flaws. Hence, the empirical performance of LinAIM should be confirmed by theoretical guarantees, and future tests may identify settings where its effectiveness is limited. Nonetheless, its design is sufficiently simple to allow substantial extensions in order to meet these theoretical criteria. In particular,

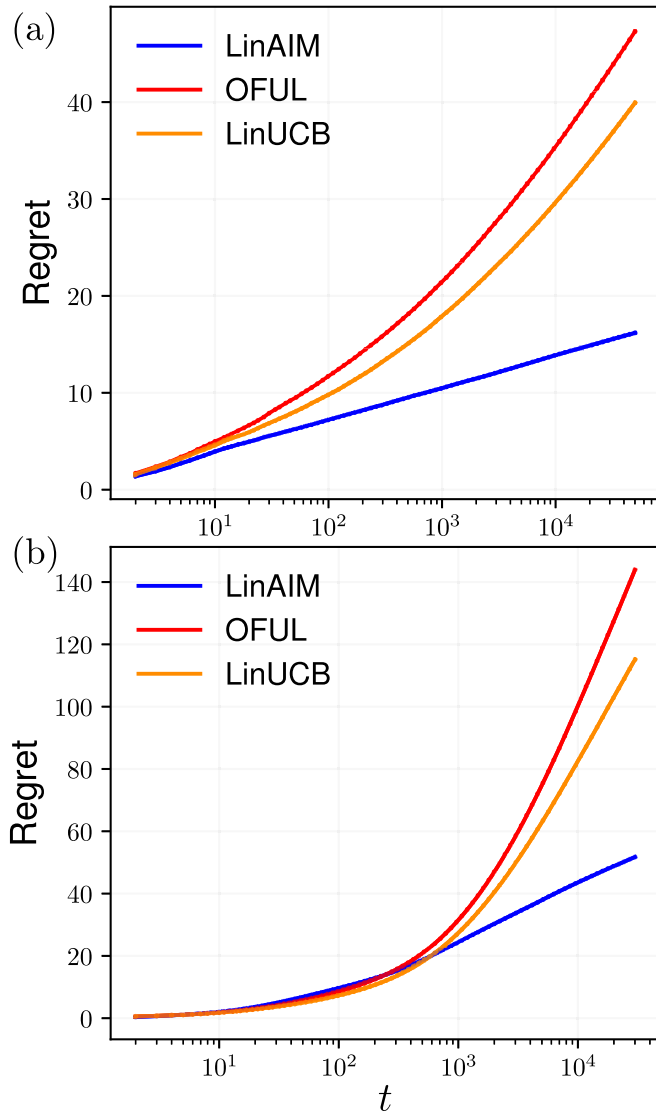


FIG. 4. Mean regret for two linear bandit settings with Gaussian rewards. In blue our algorithm and in orange and red UCB algorithms adjusted to the linear settings. (a) Bandit setting with ten arms resampled at each time from a normal distribution in $\mathbb{R}^{10}$. (b) Toy problem borrowed from Ref. [64] with two contexts. Each context (detailed in the main text) is drawn with equal probability and $\xi = 0.5$. Error bars show standard error on the mean. Details of the simulations and baseline algorithms are provided in Appendix D. To highlight the variability of the regret over individual games, a figure showing the standard deviations around the means is found in Appendix E2.

suboptimal actions could also be evaluated based on their ability to provide information about all available gaps, specifically the gaps between $\hat{\theta}_t$ and all suboptimal arms. The initial design presented in this section lays the foundation for a deeper exploration of information-based approaches for linear bandits.

V. MANY-ARMED BANDITS

A. Setting and core insight

We now consider the many-armed setting as introduced in Ref. [98]. The major differences with the classic bandit

problem are that the game's duration is finite and known in advance and the number of actions is too large for exhaustive search [Fig. 1(d)]. It underpins key applications that require multiple testing, such as those encountered in medical trials [49,50,99]. The agent must carefully define patient testing groups, balancing ethical standards and treatment costs while ensuring maximum information gain on trial efficacy. In this framework, in each round t (where $t = 1, \dots, T$), the agent selects an arm $a_t \in [K] = \{1, \dots, K\}$, where K represents the total number of arms. The chosen arm yields a stochastic reward $X_t(a_t)$, drawn from a distribution of unknown mean μ_{a_t} . We assume a Bayesian framework where the mean reward for each arm is drawn from a common prior distribution Γ with a finite support $[\mu_{\inf}, \mu_{\sup}]$. Furthermore, we assume that the reward distributions for all arms belong to the same family, parameterized solely by their means. For instance, the rewards might follow Gaussian distributions of known fixed variance σ^2 where the mean of each arm is drawn uniformly from the interval $[0, 1]$. Given a time horizon T , the agent's goal is to minimize its expected pseudoregret over T rounds in a single experiment. The pseudoregret is here given by

$$R_T = T \max_{i \in [K]} \mu_i - \sum_{t=1}^T \mathbb{E}[\mu_{a_t}]. \quad (23)$$

Ultimately, the agent aims to minimize the Bayesian expected regret,

$$\text{BR}_{T,K} = \mathbb{E}[R_T]. \quad (24)$$

Here, the expectation is taken over the prior distribution of the mean rewards, representing the agent's performance averaged over all possible mean reward configurations. The key distinction from previous settings is that the agent operates with a finite budget and has prior knowledge of the total number of rounds T . We focus on scenarios where $K > \sqrt{T}$, yielding too little time for the agent to extensively sample all the arms or fully exploit the best ones. To achieve an efficient strategy, the agent must sample only a subset of size $m \sim O(\sqrt{T})$ arms, enabling the algorithm to find a promising solution without inducing excessive exploration [98]. Beyond this scale, further exploration becomes too costly compared to leveraging the best empirical arms. Conversely, before this scale, the sampled arms have a higher probability of inducing significant regret, making early exploitation less effective [98].

Multiple strategies attain the optimal scaling of $\text{BR}_{T,K} \approx O(\sqrt{T})$ [100]. Notably, subsampled Greedy (SS-greedy) relies on an *ad hoc* selection of a subset of fixed size $O(\sqrt{T})$ to limit exploration and has been shown to be asymptotically optimal for Bernoulli rewards. While such strategies achieve optimal asymptotic scaling, they do not allow the algorithm to adapt to the rewards observed during its execution. Hence, we aim to use a physics-based approach that naturally recovers the $O(\sqrt{T})$ scaling through a functional expression, while maintaining flexibility and both short- and long-time efficiency.

B. Algorithm design

We propose an algorithm that leverages the minimization of a global functional based on underlying physics principles,

using as a proxy variable the expected upcoming regret, i.e., the regret accumulated from time t to T . Our policy works as follows: at each time step, it evaluates the expected upcoming regret for two distinct decision sequences and selects the one with the lower expected upcoming regret. The two sequences are as follows: (1) pull the current best empirical arm, denoted by M_t , until the end or (2) try a new untested arm until it eventually underperforms compared to M_t , at which point M_t is chosen. Although these strategies recommend actions for the entire remaining time, it does not mean that the algorithm will follow them in subsequent steps, as the choice between them is reassessed at each time step. Before evaluating both upcoming regrets, we need to introduce several definitions. When clear from context, we omit the dependence on t for simplicity. The number of times arm i has been pulled is denoted by t_i , and its associated empirical mean by $\hat{\mu}_i$. The number of different arms tested by time t is labeled k_t , forming the set $[k_t] = \{k_1, k_2, \dots, k_t\}$. The best empirical arm at time t is labeled M_t , with its corresponding empirical mean $\hat{\mu}_{M_t} = \max_{i \in [k_t]} \hat{\mu}_i(t)$. We also define the remaining time as $\tau = T - t$. Last, the expected maximum mean value is denoted by $\mu_* = \mathbb{E}_\Gamma[\max_{i \in [K]} \mu_i]$.

We now derive approximations for the regret in both cases; see Appendix C for further details and discussion of all approximations. For (1), i.e., fully exploit M_t for the next τ turns, the estimated upcoming expected regret is equal to

$$\tilde{R}_g = \tau(\mu_* - \hat{\mu}_{M_t}). \quad (25)$$

For (2), i.e., testing a new arm, the upcoming expected regret is estimated as

$$\begin{aligned} \tilde{R}_{\text{exp}} &= \int_{\hat{\mu}_{M_t}}^{\mu_{\text{sup}}} \Gamma(\mu)(\mu_* - \mu)\tau d\mu + \int_{\mu_{\text{inf}}}^{\hat{\mu}_{M_t}} \Gamma(\mu) \\ &\quad \times [(\mu_* - \mu)\tau_{\text{fpt}} + (\mu_* - \hat{\mu}_{M_t})(\tau - \tau_{\text{fpt}})] d\mu. \end{aligned} \quad (26)$$

The first integral captures cases where the true mean reward of the new arm is higher than $\hat{\mu}_{M_t}$, neglecting cases where the empirical mean of a promising arm never drops below $\hat{\mu}_{M_t}$. The second integral corresponds to cases where the true mean reward value of the new arm is lower than $\hat{\mu}_{M_t}$. In this case, the new arm will be pulled for an average time τ_{fpt} until it becomes suboptimal, which happens when its empirical mean falls below $\hat{\mu}_{M_t}$. This first phase yields the first term, following which the algorithm reverts to exploiting the best empirical arm, as captured by the second term.

Finally, the decision procedure evaluates

$$\begin{aligned} \Delta &= \tilde{R}_{\text{exp}} - \tilde{R}_g \\ &= \tau \int_{\hat{\mu}_{M_t}}^{\mu_{\text{sup}}} \Gamma(\mu)(\hat{\mu}_{M_t} - \mu) d\mu \\ &\quad + \int_{\mu_{\text{inf}}}^{\hat{\mu}_{M_t}} \Gamma(\mu)(\hat{\mu}_{M_t} - \mu)\tau_{\text{fpt}} d\mu \end{aligned} \quad (27)$$

and follows the strategy associated with the lowest upcoming expected regret. To get a universal and simpler expression, we employ the scaling ansatz that the typical number of steps to assess suboptimality is inversely proportional to the mean difference $\tau_{\text{fpt}} \propto (\hat{\mu}_{M_t} - \mu)^{-1}$ (refer to Appendix C for additional details).

ALGORITHM 3. Algorithm for k arms with finite horizon T .

```

Draw a first arm  $a_1$ , observe reward  $X_1$ , and update
statistics:  $\hat{\mu}_1 = X_1$ ,  $k_1 = 1$ 
for  $t = 1$  to  $T - 1$  do                                // Arm selection
     $M_t \leftarrow \underset{i \in [k_t]}{\text{argmax}} \hat{\mu}_i(t)$ 
    if  $k_t = K$  then  $a_{t+1} \leftarrow M_t$ 
    else
        Evaluate  $\Delta$  following Eq. (28)
        if  $\Delta \leq 0$  then  $a_{t+1} \leftarrow M_t$ 
        else  $a_{t+1} \leftarrow k_t + 1$ ,  $k_{t+1} \leftarrow k_t + 1$ 
        Pull  $a_{t+1}$  and observe  $X_{t+1}(a_{t+1})$ 

```

For a uniform prior over $[0, 1]$, we obtain the following expression for Δ :

$$\Delta = c\hat{\mu}_{M_t} - \tau \frac{(1 - \hat{\mu}_{M_t})^2}{2}, \quad (28)$$

where c is a free parameter, which we set to 1.

The pseudocode for the resulting algorithm, referred to as approximate expected regret minimization (ARM), is presented in Algorithm 3.

Let us briefly comment on our algorithm design. When Δ is positive, the current best empirical arm is chosen, as it involves a lower expected upcoming cost. Conversely, if Δ is negative, a new untested arm is pulled (see Fig. 5 for illustrative regret instances). Note that the algorithm behaves greedily when no more arms remain to be tested, although such cases should rarely be encountered in many-armed settings. These simplifications aim to derive a straightforward and tractable expression, while still leveraging a well-defined global quantity to minimize. Here, since the time horizon is too short for leveraging a full exploration and identification of

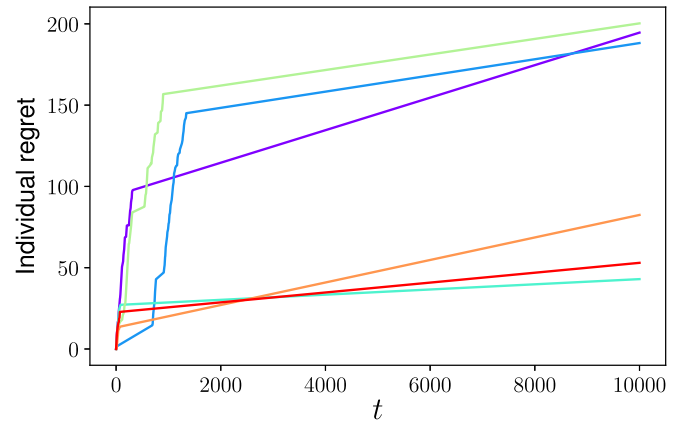


FIG. 5. Single runs of ARM for Bernoulli rewards and a uniform prior on $[0, 1]$. The regret grows until horizon T , i.e., the stopping time, is reached $T = 10000$ ($K = 500$ and $c = 1$). For each execution, the agent stops its exploration phase once a sufficiently promising reward has been obtained. Therefore, this “stopping time” may occur at different times. In the first “exploitation phase,” the regret shows a linear slope as the algorithm consistently selects the same reward, whereas in the exploration phase, it exhibits a steeper, stair-step pattern indicating varied reward selection. Details of the algorithm and simulations are provided in Appendix D.

the best arm, the need to acquire more information must be significantly limited in the decision process. Therefore, here the relevant information lies in the time required to assess that a new arm is suboptimal and is quantified by τ_{fpt} .

C. Results

To assess the efficiency of ARM (Algorithm 3), its average regret is compared with the state-of-the-art SS-epsilon greedy algorithm (see Ref. [98] and Appendix D 2 for implementation details). Notably, SS-epsilon greedy has been shown to achieve an optimal regret scaling of $O(\sqrt{T})$ for $K \gg \sqrt{T}$. Figure 6(a) shows the regret in the many-armed bandit setting with a finite horizon and Bernoulli rewards, where our algorithm outperforms the SS-greedy algorithm. This is also confirmed for several horizon durations [Fig. 6(b)]. This numerical result outlines that the optimal balance between exploration and exploitation is already embedded in the underlying physical structure of our algorithm. In contrast, SS-greedy manually imposes a first exploration phase of constant duration (namely, $\sqrt{T}$) to ensure optimal scaling, followed by a greedy strategy, leading to disjoint exploration-exploitation phases. This is exemplified by ARM displaying a smoother initial regret curve compared to SS-greedy. ARM alternates between exploration and exploitation until a sufficiently promising arm is found, adapting its exploration amount around the fixed $\sqrt{T}$ ratio if a sufficiently promising arm is found earlier or later [101]. Notably, in Fig. 6(a), the selected arm is found later by ARM than by the SS-greedy policy but outperforms the one found by SS-greedy on average, as indicated by the distinct final slopes of the regret closer to the horizon.

The empirically observed optimality of ARM is supported by a heuristic scaling argument for when the transition between exploration and exploitation likely occurs (see Appendix C for additional elements). Briefly, the expected value of Δ [Eq. (28)] crosses over from positive to negative values for $\tau(1 - \hat{\mu}_{M_t}) \sim O(1)$. Additionally, the typical value of $\hat{\mu}_{M_t}$ scales with t as $(1 - \hat{\mu}_{M_t}) \sim 1/t$ since it is inversely proportional to the number of tested arms, which is in turn proportional to t in the exploration phase (before the crossover from positive to negative Δ). Combining these scales implies that $t \sim O(\sqrt{T})$, thereby explaining the optimal scaling of ARM's regret. Note, however, that this argument does not constitute a rigorous proof, which would need to account for the influence of rare events, such as failing to identify an arm for exploitation within $O(\sqrt{T})$ time.

We conclude this section by several remarks. First, ARM relies on its ability to combine its knowledge of the remaining time and the available information on Γ to optimize its decision process. Notably, Eq. (26) applies to more general prior distributions. If the prior is not provided to the agent, ARM could be adapted to add an empirical evaluation of Γ , in particular close to its upper bound. Similar to SS-greedy, ARM exploits by selecting the arm with the best empirical mean, which still allows identifying a sufficiently efficient arm within the subset of tested arms. To enhance the exploitation component and leverage prior knowledge, one may consider selecting the arm with the highest probability of being optimal, rather than the highest posterior mean. This

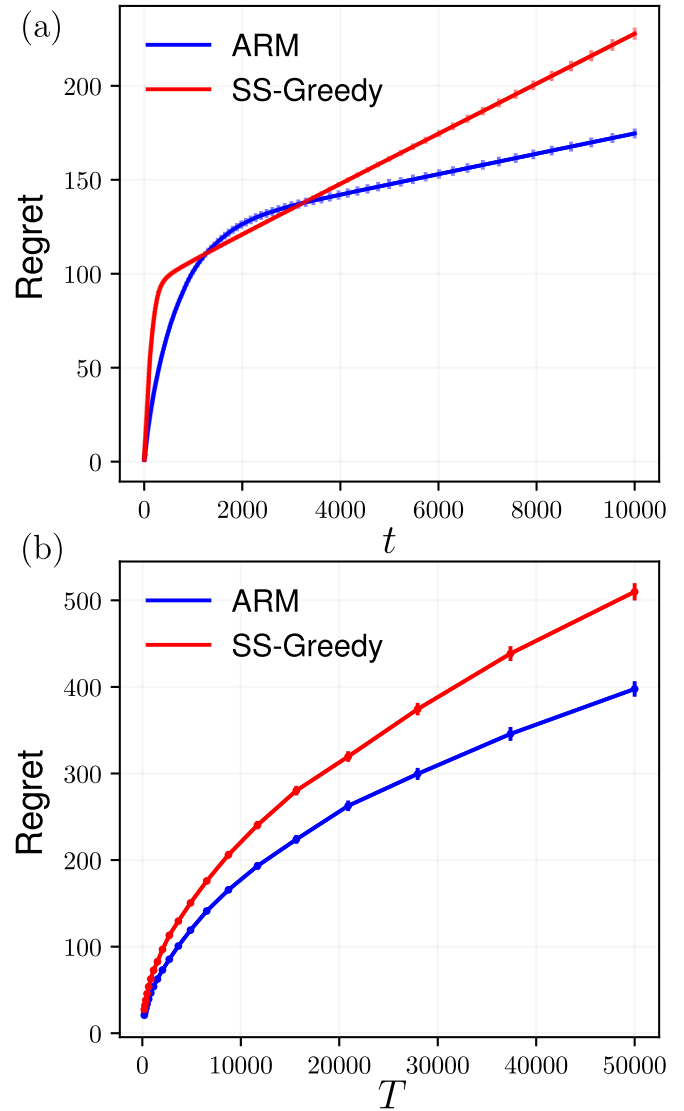


FIG. 6. Mean regret for the many-armed bandit problem with Bernoulli rewards and a uniform prior on $[0, 1]$. (a) The regret growth of our algorithm (ARM), in blue, is observed until horizon T , i.e., the stopping time, is reached $T = 10\,000$ ($K = 5000 \gg \sqrt{T}$ and $c = 1$). The regret is averaged over 2000 runs and standard errors are indicated (see Appendixes D and D 2 for details on the numerical settings). Its performance is compared to the ss-greedy algorithm, in red. The final slopes of both algorithms differ, indicating that the selected arm for most of the exploitation is less efficient than the one found for our algorithm. Additionally, our algorithm presents a slower initial slope because exploration and exploitation are already mixed. (b) Mean regret performance when at the stopping time (when the horizon is reached), for distinct games with varying horizon length ($K = T \gg \sqrt{T}$ and $c = 1$). The regret is averaged over 2000 runs with error bars indicating standard error on the mean. Details of the simulations and the SS-greedy algorithm are provided in Appendix D. To highlight the variability of the regret over individual games, a figure showing the standard deviations around the means is found in Appendix E 3.

slight difference is crucial, as the algorithm might otherwise fall into exploiting arms with better averages but a higher probability of being suboptimal compared to other explored arms. To implement such a strategy, one could replace $\hat{\mu}_{M_t}$ in

Eq. (28) by the expected mean reached by playing this alternative strategy given the current tested arms. However, such an alternative remains more complex than using the empirical mean as we showed already provides efficient results. This approach could be considered for future developments. This flexibility should favor the extension and adaptability of ARM to more complex priors and bandit problems.

VI. CONCLUSION

This study presents a generalization of the information-maximization principle and physics-based representations to a varied selection of structured bandit games. Our study addresses three distinct problems within the multiarmed bandit domain: pure exploration in the form of the explore- m problem; contextual problems, exemplified by linear bandits; and many-armed bandits. Each of these settings represents a unique facet of the broad field of bandit games, and each requires a tailored approach to account for their underlying structures and achieve the right balance between exploration and exploitation. To this end, for each setting, we focused on a specific observable from which we extracted a simplified and analytical functional at the core of the decision scheme. Numerical experiments in the paradigmatic versions of these games showed the robustness and efficiency of our method. Further research should focus on deriving rigorous theoretical bounds tailored to each bandit setting. Our method goes beyond classical confidence intervals for individual arms by leveraging a functional representation to guide decision-making. This core architecture is designed to make our method more adaptable to the diverse variants that may be encountered in real-world scenarios. Each algorithm examined above could be adapted to address more intricate variations of the corresponding problems, especially when dealing with reward or prior distributions exhibiting power-law tails, which induce extreme values. Future work will focus on searching for demonstrations of optimality (such as the ones derived for AIM) and systematic approaches to devise information-based approaches to decision-making problems.

ACKNOWLEDGMENTS

We acknowledge the help of the HPC Core Facility of the Institut Pasteur for this work. This study was funded by the Agence Nationale de la Recherche Grants No. ANR-20-CE45-0021 and No. ANR-23-CE23-0034 (microBrain on Edge), the programmes managed by Agence Nationale de la Recherche under the France 2030 programme, ANR-23-IACL-0008 (PRAIRIE 3IA Institute), ANR-23-IAHU-0003 (IHU ReConnect), ANR-23-IAIU-0010 (IHU ICE), ANR-23-RHUS-0011 (ReBone), and ANR-24-RRII-0002 operated by the Inria Quadrant Programme. The Per Fumum sponsorship also supported this work. We finally acknowledge the help of the Pasteur-Roux-Cantarini Fellowship for this work.

DATA AVAILABILITY

The code and simulated data that support the findings of this article are openly available [102].

APPENDIX A: DETAILS OF EXPLORE- m DESIGN

We provide all derivation details leading to PacAIM expression given in the main text.

1. Link between information maximization and probability maximization

Here, we briefly comment on the decision to switch from entropy increments to probability increments in the arm selection process of our PacAIM algorithm. The associated entropy for all arms, effectively well separated by θ_b , is given by

$$S_{\text{top}} = -P_{\text{top}} \ln P_{\text{top}} - (1 - P_{\text{top}}) \ln(1 - P_{\text{top}}), \quad (\text{A1})$$

with $P_{\text{top}} = \prod_{i \in \mathcal{M}_t} \mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2})$. The information gain relative to arm $i \in \mathcal{M}_t$ thus reads

$$\Delta S_{\text{top}} = -\Delta P_{\text{top}} [\ln(P_{\text{top}}) - \ln(1 - P_{\text{top}})]. \quad (\text{A2})$$

For $P_{\text{top}} > 1/2$, the sign of ΔS_{top} is given by ΔP_{top} , meaning that minimizing the entropy resumes to selecting the arm maximizing ΔP_{top} . At short times, one may face events with $P_{\text{top}} < 1/2$. For such cases, direct entropy minimization will be inaccurate since Eq. (A2) ignores that arms may switch from $\mathcal{M}_t$ to $\mathcal{D}_t$, resulting in additional variations in P_{top} not caught in Eq. (A2). We thus assume to minimize to select arms maximizing $\Delta P_{\text{top}} |\ln(P_{\text{top}}) - \ln(1 - P_{\text{top}})|$ increment. The same argument can be drawn for P_{bot} .

2. Details on increment derivation

We start by reminding P_{top} expression:

$$P_{\text{top}} = \prod_{i \in \mathcal{M}_t} \mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2}). \quad (\text{A3})$$

By leveraging the arms' independence, the product enables a straightforward computation of the increments along each arm:

$$\Delta_{i \in \mathcal{M}_t} P_{\text{top}} = \frac{\Delta \mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2})}{\mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2})} P_{\text{top}}. \quad (\text{A4})$$

Since all gradients can be factorized by $P_{\text{bot}} P_{\text{top}}$, the decision criteria reduce to a comparison of the relative gradients of the cumulative distribution associated with each arm. We finally conclude by computing the gradient of the cumulative probability associated with the arm i . It decomposes into partial derivatives along the empirical mean $\hat{\mu}_i$ and the number of draws t_i . We first notice that the partial derivative along $\hat{\mu}_i$ cancels by symmetry. Then, we consider the derivative along the number of draws.

For Gaussian rewards with unit variance, we have

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_t}(\mu_i \geq \theta_b - \frac{\epsilon}{2}) &= 1 - \frac{1}{2} \left[1 + \operatorname{erf} \left(\sqrt{\frac{t_i}{2}} \left[\theta_b - \frac{\epsilon}{2} - \hat{\mu}_i \right] \right) \right] \\ &= \frac{1}{2} \left[1 + \operatorname{erf} \left(\sqrt{\frac{t_i}{2}} \left[\hat{\mu}_i - \theta_b + \frac{\epsilon}{2} \right] \right) \right] \end{aligned} \quad (\text{A5})$$

and

$$\begin{aligned} \Delta \mathbb{P}_{\mathcal{H}_t} \left(\mu_i \geq \theta_b - \frac{\epsilon}{2} \right) &= \frac{\partial}{\partial t_i} \mathbb{P}_{\mathcal{H}_t} \left(\mu_i \geq \theta_b - \frac{\epsilon}{2} \right) \\ &= \frac{\hat{\mu}_i - \theta_b + \frac{\epsilon}{2}}{2\sqrt{2\pi}t_i} \exp \left(-t_i \frac{(\theta_b - \frac{\epsilon}{2} - \hat{\mu}_i)^2}{2} \right). \end{aligned} \quad (\text{A6})$$

Replicating this reasoning for P_{bot} leads to

$$\begin{aligned} \Delta \mathbb{P}_{\mathcal{H}_t} \left(\mu_i \leq \theta_b + \frac{\epsilon}{2} \right) &= \frac{\partial}{\partial t_i} \mathbb{P}_{\mathcal{H}_t} \left(\mu_i \geq \theta_b - \frac{\epsilon}{2} \right) \\ &= \frac{\partial}{\partial t_i} \left(\frac{1}{2} \left[1 + \operatorname{erf} \left(\sqrt{\frac{t_i}{2}} \left[\theta_b + \frac{\epsilon}{2} - \hat{\mu}_i \right] \right) \right] \right) \\ &= \frac{\theta_b + \frac{\epsilon}{2} - \hat{\mu}_i}{2\sqrt{2\pi}t_i} \exp \left(-t_i \frac{(\theta_b + \frac{\epsilon}{2} - \hat{\mu}_i)^2}{2} \right). \end{aligned} \quad (\text{A7})$$

3. Details on θ_b derivation

Here, we provide additional details on the derivation steps leading to θ_b expression in the main text. Our goal is to derive a tractable expression of an appropriate cutoff between the subset of arms with the current m highest empirical means, $\mathcal{M}_t$, with a minimal empirical min $\hat{\mu}_{\text{top}} = \max(\{\hat{\mu}_{\mathcal{D}_t}\})$ and the remaining arms of maximal mean $\mathcal{M}_t$, $\hat{\mu}_{\text{bot}} = \min(\{\hat{\mu}_{\mathcal{D}_t}\})$.

An optimal θ_b value should permit the algorithms to successfully stop in a minimal amount of samples denoted by τ in the following. For a given set of observed empirical rewards $\{\mu_{\mathcal{M}_t}\}$, $\{\mu_{\mathcal{D}_t}\}$ and neglecting their future variations, an oracle strategy will verify

$$\begin{aligned} \theta_{b,o} = \operatorname{argmin}_{\theta \in [\hat{\mu}_{\text{bot}}, \hat{\mu}_{\text{top}}]} \left[\tau = \sum_i \tau_i \text{ s.t. } \prod_{i \in \mathcal{D}_t} \mathbb{P} \left(\mu_i \leq \theta + \frac{\epsilon}{2} | \hat{\mu}_i, \tau_i \right) \right. \\ \left. \times \prod_{i \in \mathcal{M}_t} \mathbb{P} \left(\mu_i \geq \theta - \frac{\epsilon}{2} | \hat{\mu}_i, \tau_i \right) > 1 - \delta \right]. \end{aligned} \quad (\text{A8})$$

Here, $\theta_{b,o}$ is the optimal value that minimizes τ while satisfying the stopping criteria and $\{\tau_1, \dots, \tau_n\}$ are the local times of each arm at the stopping time. Probabilities in Eq. (A8) depend on the hypothetical number of samples for each arm, τ_i , which differs by the current individual arm time t_i for P_{top} and P_{bot} . At short times, $\tau_i \gg t_i$ because it takes a certain number of samples before meeting the conditions in Eq. (A8). However, by selecting a θ_b close to $\theta_{b,o}$, the algorithm should adjust its strategy to match t_i with τ_i as closely as possible.

Unfortunately, the expression for $\theta_{b,o}$ becomes intractable with more than two arms. To derive a simplified expression, we approximate $\mathcal{M}_t$ as a fictive set of m arms, all with the same empirical mean, $\hat{\mu}_{\text{top}}$, and the same number of pulls, $t_{\text{top}} = T_{\text{top}}/m$. Because all these fictive arms share the same empirical mean and pull count, the posterior distribution of their minimal mean value can be approximated by a Gumbel

law [76] parametrized by $(\hat{\mu}_{\text{top}}, t_{\text{top}}, m)$.

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_t} \left(\min_{i \in \{1, \dots, m\}} (\mu_i | \hat{\mu}_i = \hat{\mu}_{\text{top}}, t_i = t_{\text{top}}) \geq \theta \right) \\ = \exp[-\exp(\sqrt{2 \ln(m)} [\sqrt{t_{\text{top}}}(\theta - \hat{\mu}_{\text{top}}) + \sqrt{2 \ln(m)}])]. \end{aligned} \quad (\text{A9})$$

This heuristic approximation is particularly relevant for large set values m , which are the most complicated cases for which relying on an efficient θ_b value matters the most. This approximation is discussed in greater detail in Appendix A.4.

By replicating the previous scheme for $\mathcal{D}_t$, we obtain a Gumbel law parametrized by $(\hat{\mu}_{\text{bot}}, t_{\text{bot}}, d = n - m)$ with a cumulative distribution reading

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_t} \left(\max_{i \in \{1, \dots, d\}} (\mu_i | \hat{\mu}_i = \hat{\mu}_{\text{bot}}, t_i = t_{\text{bot}}) \leq \theta \right) \\ = \exp[-\exp(-\sqrt{2 \ln(d)} [\sqrt{t_{\text{bot}}}(\theta - \hat{\mu}_{\text{bot}}) - \sqrt{2 \ln(d)}])]. \end{aligned} \quad (\text{A10})$$

Let us note that the centering parameters $\sqrt{2 \ln(m)}$ and $\sqrt{2 \ln(d)}$ result from the Gaussian property of the reward distribution. Hence, we define θ_b as simultaneously verifying

$$\mathbb{P}_{\mathcal{H}_t} \left(\min_{i \in \{1, \dots, m\}} (\mu_i | \hat{\mu}_i = \hat{\mu}_{\text{top}}, t_i = t_{\text{top}}) \geq \theta_b - \frac{\epsilon}{2} \right) = 1 - \frac{\delta}{2} \quad (\text{A11})$$

and

$$\mathbb{P}_{\mathcal{H}_t} \left(\max_{i \in \{1, \dots, d\}} (\mu_i | \hat{\mu}_i = \hat{\mu}_{\text{bot}}, t_i = t_{\text{bot}}) \leq \theta_b + \frac{\epsilon}{2} \right) = 1 - \frac{\delta}{2}. \quad (\text{A12})$$

These events include all cases where the fictive subset is not ϵ -optimal, upper bounding the identification error probability by δ . By injecting Eqs. (A9) and (A10) within Eqs. (A11) and (A12), one obtains

$$\theta_b = \frac{\epsilon}{2} + \hat{\mu}_{\text{top}} - \frac{1}{\sqrt{t_{\text{top}}}} \left[\sqrt{2 \ln(m)} - \frac{\ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(m)}} \right] \quad (\text{A13})$$

and

$$\theta_b = -\frac{\epsilon}{2} + \hat{\mu}_{\text{bot}} + \frac{1}{\sqrt{t_{\text{bot}}}} \left[\sqrt{2 \ln(d)} - \frac{\ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(d)}} \right]. \quad (\text{A14})$$

By combining Eqs. (A13) and (A14), one obtains

$$\begin{aligned} \epsilon + \hat{\mu}_{\text{top}} - \hat{\mu}_{\text{bot}} = \frac{1}{\sqrt{t_{\text{top}}}} \left[\sqrt{2 \ln(m)} - \frac{\ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(m)}} \right] \\ + \frac{1}{\sqrt{t_{\text{bot}}}} \left[\sqrt{2 \ln(d)} - \frac{\ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(d)}} \right]. \end{aligned} \quad (\text{A15})$$

We finally look for t_{top} and t_{bot} verifying Eq. (A15) but leading to a minimal amount of time $mt_{\text{top}} + dt_{\text{bot}}$. By labeling

$$\begin{aligned} c_{\text{top}} &= \left[\sqrt{2 \ln(m)} - \frac{\ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(m)}} \right], \\ c_{\text{bot}} &= \left[\sqrt{2 \ln(d)} - \frac{\ln(-\ln(1 - \frac{\delta}{2}))}{\sqrt{2 \ln(d)}} \right], \\ M &= \epsilon + \hat{\mu}_{\text{top}} - \hat{\mu}_{\text{bot}}, \end{aligned} \quad (\text{A16})$$

we obtain

$$t_{\text{top}} = \left[\frac{c_{\text{top}}}{M - \frac{c_{\text{bot}}}{\sqrt{t_{\text{bot}}}}} \right]^2 \quad (\text{A17})$$

and

$$t_{\text{bot}} = \frac{c_{\text{bot}}}{M^2} + \left(\frac{c_{\text{bot}}^2 c_{\text{top}}^4 m^2}{M^6 d^2} \right)^{\frac{1}{3}} + \frac{2M^2 d}{c_{\text{top}}^2 m} \left(\frac{c_{\text{bot}}^2 c_{\text{top}}^4 m^2}{M^6 d^2} \right)^{\frac{2}{3}}, \quad (\text{A18})$$

which simplifies into

$$t_{\text{bot}} = \frac{c_{\text{bot}}}{M^2} + \left(\frac{c_{\text{bot}}^2 c_{\text{top}}^4 m^2}{M^6 d^2} \right)^{\frac{1}{3}} \left(1 + \frac{2d^{\frac{2}{3}} c_{\text{bot}}^{\frac{2}{3}}}{m^{\frac{1}{3}} c_{\text{top}}^{\frac{2}{3}}} \right). \quad (\text{A19})$$

Inserting Eq. (A19) in Eq. (A14) gives the desired analytical expression of θ_b :

$$\begin{aligned} \theta_b &= -\frac{\epsilon}{2} + \hat{\mu}_{\text{bot}} + c_{\text{bot}} \\ &\times \left[\frac{c_{\text{bot}}}{M^2} + \left(\frac{c_{\text{bot}}^2 c_{\text{top}}^4 m^2}{M^6 d^2} \right)^{\frac{1}{3}} + \frac{2M^2 d}{c_{\text{top}}^2 m} \left(\frac{c_{\text{bot}}^2 c_{\text{top}}^4 m^2}{M^6 d^2} \right)^{\frac{2}{3}} \right]^{-\frac{1}{2}}. \end{aligned} \quad (\text{A20})$$

Note that our approximation scheme is mainly independent of the reward distribution, as it only involves c_{bot} and c_{top} expressions. To extend our results to various reward distributions, one simply needs to adjust the Gumbel law by providing the corresponding centering parameters (here, $\sqrt{2 \ln(m)}$ for Gaussian rewards). Leveraging the scaling proposed in Ref. [76] for any distribution exhibiting a large x tail $p(x) \sim \exp(-x^\gamma)$, the adjusted Gumbel law distributions read

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_t} \left(\min_{i \in [1, \dots, m]} (\mu_i | \hat{\mu}_i = \hat{\mu}_{\text{top}}, t_i = t_{\text{top}}) \geq \theta \right) \\ = \exp \left[-\exp \left(\gamma \ln(m)^{1-\frac{1}{\gamma}} \left[\sqrt{t_{\text{top}}} (\theta - \hat{\mu}_{\text{top}}) + \ln(m)^{\frac{1}{\gamma}} \right] \right) \right] \end{aligned} \quad (\text{A21})$$

and

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_t} \left(\max_{i \in [1, \dots, d]} (\mu_i | \hat{\mu}_i = \hat{\mu}_{\text{bot}}, t_i = t_{\text{bot}}) \leq \theta \right) \\ = \exp \left[-\exp \left(\gamma \ln(d)^{1-\frac{1}{\gamma}} \left[\sqrt{t_{\text{bot}}} (\theta - \hat{\mu}_{\text{bot}}) + \ln(d)^{\frac{1}{\gamma}} \right] \right) \right], \end{aligned} \quad (\text{A22})$$

only affecting c_{bot} and c_{top} expressions. Still, keeping the Gaussian approximation within θ_b should remain reasonable for non-heavy-tailed reward distributions.

4. Note on the Gumbel law approximation

In this section, we briefly discuss the Gumbel law approximation leading to our expression for θ_b . We emphasize that the approximation is a heuristic to derive a simple evaluation criterion, rather than an approach with established convergence guarantees under asymptotic conditions. This is because the arms do not share the same mean and, logically, require a distinct number of draws using an efficient algorithm. Consequently, obtaining an asymptotically exact approximation of the maximum becomes analytically intractable as far as we know.

To be more precise, t_{top} and t_{bot} are derived as the optimal numbers of draws for each set, assuming that all arms' means are equal to either $\hat{\mu}_{\text{top}}$ or $\hat{\mu}_{\text{bot}}$, a condition that is neither achieved nor targeted by our algorithm or any optimal policy. For example, the worst and best arms will be drawn a negligible number of times whenever their means are not close to $\hat{\mu}_{\text{top}}$ or $\hat{\mu}_{\text{bot}}$, respectively. This induces a distribution that is more spread out than the Gumbel distribution (see Fig. 9). Furthermore, we assume that the cumulative distribution functions of the Gumbel distributions should both be equal to $1 - \delta/2$ at θ_b , which is often not achieved for all mean priors, also preventing convergence [see Fig. 9(c) for an illustrative example].

Nonetheless, we emphasize that this does not undermine the algorithm's efficiency, as the stopping criteria and increment computations are precise. In other words, we ensure that our algorithm evaluates its exact stopping point and increments at θ_b . While θ_b may not be asymptotically optimal for all mean priors, it facilitates a simpler stopping criterion compared to upper confidence bound solutions. Looking forward, future research could explore solutions based on the integration of increments between $[\hat{\mu}_{\text{bot}}, \hat{\mu}_{\text{top}}]$.

APPENDIX B: DETAILS OF LinAIM FOR LINEAR BANDITS

Here, we provide derivations and simplification details regarding our algorithm for the linear bandit setting.

1. Note on posterior distribution associated with θ_* value

We first seek to derive the posterior probability distribution of θ_* value given the history $\mathcal{H}_t$, which consists of past pairs of chosen arms and their corresponding rewards (A_n, X_n) for $n \in [0, t]$. We also rely on a Gaussian prior for θ_* value with an L^2 norm:

$$p(\theta_* = \theta) = \mathcal{N}^{-1} \exp \left(-\frac{\lambda}{2} \sum_i^d \theta_i^2 \right), \quad (\text{B1})$$

where d is the arm space dimension. Choosing $\lambda > 0$ serves two purposes: first, it ensures that the covariance matrix of the associated posterior distribution is invertible even when no arms have been drawn. Second, it focuses the inference on more realistic optimal values by expecting a mean reward close to the unit norm. According to Bayes's theorem, the

posterior distribution reads

$$\begin{aligned}
p(\theta_* = \theta | \mathcal{H}_t) &= p(\mathcal{H}_t | \theta_*) p(\theta_* = \theta) p(\mathcal{H}_t)^{-1} \\
&= p(\mathcal{H}_t)^{-1} \frac{\mathcal{N}^{-1} e^{-\frac{1}{2} \sum_{n=0}^t \theta_n^2}}{(2\pi)^{t/2}} \exp\left(-\frac{1}{2} \sum_{n=0}^t [\theta' A_n - X_n]^2\right) \\
&= p(\mathcal{H}_t)^{-1} \frac{\mathcal{N}^{-1} e^{-\frac{1}{2} \sum_{n=0}^t X_n^2}}{(2\pi)^{t/2}} \times \dots \\
&\quad \times \exp\left(\theta' \sum_{n=0}^t X_n A_n - \frac{1}{2} \theta' \left[\sum_{n=0}^t A_n A_n' + \lambda I_d \right] \theta\right) \\
&= \frac{\exp\left(-\frac{1}{2} (\theta - \hat{\theta}_t)' B_t (\theta - \hat{\theta}_t)\right)}{\sqrt{(2\pi)^d \det B_t^{-1}}}, \tag{B2}
\end{aligned}$$

where we defined $B_t = \lambda I_d + \sum_{n=0}^t A_n A_n'$ and $\hat{\theta}_t = B_t^{-1} \sum_{n=0}^t X_n A_n$. The final step relies on the distribution being normalized. Equation (B2) holds for Gaussian rewards and one recognizes a multivariate Gaussian distribution. Equation (B2) forms the basis for the entropy calculations that underpin our algorithm.

2. Details of entropy increments for linear bandits LinAIM

Here, we outline the key steps for computing the entropy increments discussed in the main text.

a. Entropy assigned to the best available empirical arm

We first consider the best available empirical arm at time t defined as $A_{M_t} = \max_{a \in \mathcal{A}_t} \langle \theta_*, a \rangle$. We assign to this arm the information gain of θ_* value whose posterior is given by Eq. (B2). Because Eq. (B2) distribution is a Gaussian multivariate, the associated entropy reads

$$\begin{aligned}
S &= - \int_{\mathbb{R}^d} p(\theta_* = \theta | \mathcal{H}_t) \ln p(\theta_* = \theta | \mathcal{H}_t) d\theta \\
&= \frac{1}{2} \ln(\det B_t^{-1}) + \frac{d}{2} [1 + \ln(2\pi)]. \tag{B3}
\end{aligned}$$

The resulting increment hence reads

$$\begin{aligned}
\Delta S &= -\frac{1}{2} \ln \left[\frac{\det(B_t + A_{M_t} A_{M_t}')}{\det(B_t)} \right] \\
&= -\frac{1}{2} \ln \left[\frac{(1 + A_{M_t}' B_t^{-1} A_{M_t}) \det(B_t)}{\det(B_t)} \right] \\
&= -\frac{1}{2} \ln(1 + A_{M_t}' B_t^{-1} A_{M_t}), \tag{B4}
\end{aligned}$$

where we used the matrix determinant lemma with B_t^{-1} and $A_{M_t} A_{M_t}'$.

b. Entropy assigned to the empirical suboptimal arms

Next, we consider the entropy assigned to the remaining suboptimal actions of $\mathcal{A}_t$. For each arm, denoted by A_t , we assign the information gain regarding the event $\theta_*'(A_{M_t} - A_t) >$

0 of probability $C_{A_{M_t}, A_t}$:

$$C_{A_{M_t}, A_t} = \dots \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\hat{\theta}_t' [A_{M_t} - A_t]}{\sqrt{2 [A_{M_t} - A_t]' B_t^{-1} [A_{M_t} - A_t]}} \right) \right], \tag{B5}$$

with its associated entropy

$$\begin{aligned}
S_{C, A_t} &= -C_{A_{M_t}, A_t} \ln C_{A_{M_t}, A_t} \dots \\
&\quad - (1 - C_{A_{M_t}, A_t}) \ln(1 - C_{A_{M_t}, A_t}). \tag{B6}
\end{aligned}$$

To get a simplified expression, we made a first-order expansion of $C_{A_{M_t}, A_t}$ and of its entropy. As for AIM approaches, we only focus on the asymptotic regime where the probability of a seemingly suboptimal arm being optimal is a rare event, in other words when $C_{A_{M_t}, A_t} \approx 1$. Surprisingly, this approximation is still accurate enough outside this asymptotic regime to provide a high-performance decision scheme. We first make a first-order expansion in the regime $C_{A_{M_t}, A_t} \approx 1$. Denoting the inner term of the error function in Eq. (B5) as $\frac{\delta}{\sqrt{2\Sigma}}$ with $\delta = \hat{\theta}_t' [A_{M_t} - A_t]$ and $\Sigma = [A_{M_t} - A_t]' B_t^{-1} [A_{M_t} - A_t]$ leads to

$$C_{A_{M_t}, A_t} = \frac{e^{-\frac{\delta^2}{2\Sigma}}}{2 \frac{\delta}{\sqrt{2\Sigma}} \sqrt{\pi}} (1 + o(1)) \tag{B7}$$

and

$$\tilde{S}_{C, A_t} \approx \frac{\delta}{\sqrt{\Sigma}} \frac{e^{-\frac{\delta^2}{2\Sigma}}}{2\sqrt{\pi}}. \tag{B8}$$

Applying Eq. (B4) to S for A_t and multiplying it with Eq. (B8) viewed as a weight leads to the final expression given in the main text.

APPENDIX C: DETAILS OF ARM FOR THE MANY-ARMED SETTINGS

Here, we recapitulate all the steps leading to the analytical expression constitutive of our algorithm in the many-arm setting. We stress that it involves considerably simplifying our functional expression to get the final form of ARM. We will also evoke extensions to more general prior distributions.

We start by commenting on the partition scheme and their related approximations leading to both upcoming regret expressions given in the main text. The regret induced by solely selecting the currently better empirical arm is given by

$$\begin{aligned}
\tilde{R}_g &= \tau \left(\mu_* - \int_{\mu_{\inf}}^{\mu_{\sup}} \mu p(\mu_{M_t} = \mu | \hat{\mu}_{M_t}, t_{M_t}, \Gamma) d\mu \right) \\
&\approx \tau (\mu_* - \hat{\mu}_{M_t}), \tag{C1}
\end{aligned}$$

where τ denotes the remaining time and the integral spans over the mean reward of the current best empirical mean based on its history. To be more accurate, the resulting regret of a policy reevaluating the current best empirical arm at each time step should also account for the possibility that the best empirical arm may change over time. Yet, approximating this policy by the simpler strategy given by Eq. (C1) significantly simplifies the derivation without introducing undesirable behavior. We also approximate the mean posterior value by the

empirical mean in Eq. (C1), which provides a greedy strategy to explore the most promising arm. Next, we derive the second upcoming regret expression given in the main text. We make a partition based on how the mean reward of the newly explored arm behaves compared to $\hat{\mu}_{M_i}$:

$$\begin{aligned} \tilde{R}_{\text{exp}} = & \tau \int_{\hat{\mu}_{M_i}}^{\mu_{\text{sup}}} \Gamma(\mu)(\mu_* - \mu) d\mu + \int_{\mu_{\text{inf}}}^{\hat{\mu}_{M_i}} \Gamma(\mu)[(\mu_* - \mu)\tau_{\text{fpt}} \\ & + (\mu_* - \hat{\mu}_{M_i})(\tau - \tau_{\text{fpt}})] d\mu. \end{aligned} \quad (\text{C2})$$

The first term approximates cases where the new arm is better than $\hat{\mu}_{M_i}$. Here, we neglect cases where the new arm's mean falls below $\hat{\mu}_{M_i}$. The second term accounts for cases where the new arm mean is lower than $\hat{\mu}_{M_i}$. Our policy will continue pulling such unfavorable arms until their mean reward falls below $\hat{\mu}_{M_i}$, an event occurring on average after τ_{fpt} time steps. Finally, for consistency with Eq. (C1), we approximate the regret induced by solely selecting the current best empirical arm using $\hat{\mu}_{M_i}$. An efficient policy should avoid allocating exploration time to previously explored arms that already appear to be suboptimal, even if there is a finite probability that they may be the better choice. Given the large number of arms, our policy will either select arms with empirical means greater than or equal to $\hat{\mu}_{M_i}$ or explore an untested arm. Indeed, if the current best empirical mean starts to decline, there will always be a nearby arm with a similar empirical mean. Iterating this process, it becomes clear that the algorithm will end up finding an arm with a mean close to, if not better than, $\hat{\mu}_{M_i}$. This observation justifies that our greedy policy can be effectively approximated by $\hat{\mu}_{M_i}$. Therefore, we assume that cases where a better arm falls below $\hat{\mu}_{M_i}$ can be neglected, as such an arm will no longer impact the regret. Such probability may also be derived but does not seem to improve the algorithm's efficiency.

We now present the last steps to get a tractable expression, i.e., obtaining τ_{fpt} scaling. We label by i the new tested arm:

$$\tau_{\text{fpt}} = \min(\mathbb{E}[\min(s > t \mid \hat{\mu}_{M_i} > \hat{\mu}_i(s)) \mid \mu_i = \mu], \tau), \quad (\text{C3})$$

where arm i is drawn at each time until either $\hat{\mu}_i(s)$ falls below $\hat{\mu}_{M_i}$ or the time horizon is reached. Since the successive rewards are independent from each other, τ_{fpt} can be modeled by the first time where $\sum_i^n X_i < n\hat{\mu}_{M_i}$, with X_i being independent identically distributed according to the reward distribution parametrized by μ . It thus resumes as the mean first-passage time of a biased random walk dropping below $\hat{\mu}_{M_i}$. For an asymmetric jump distribution, τ_{fpt} may be deduced from a generalized version of the Sparre-Andersen theorem [103–105]. By defining $q_+(n = \text{Prob}[x_n \geq 0, x_{n-1} \geq 0, \dots \mid x_0 = 0])$ and its associated generative series $\tilde{q}_+(s) = \sum_s q_+(n)s^n$, it states that

$$\tilde{q}_+(s) = \exp \left[\sum_1^\infty \frac{s^n}{n} \text{Prob} \left(\sum_i^n X_i \geq \hat{\mu}_{M_i} n \right) \right]. \quad (\text{C4})$$

By noticing

$$\begin{aligned} \tau_{\text{fpt}} = & \sum_{n=1}^\infty n[q_+(n-1) - q_+(n)] \\ = & \sum_{n=1}^\infty q_+(n-1) + \sum_{n=1}^\infty (n-1)q_+(n-1) - \sum_{n=1}^\infty nq_+(n) \\ = & \tilde{q}_+(1), \end{aligned} \quad (\text{C5})$$

we obtain using Eq. (C4)

$$\tau_{\text{fpt}} = \exp \left[\sum_1^\infty \frac{1}{n} \text{Prob} \left(\sum_i^n X_i \geq \hat{\mu}_{M_i} n \right) \right]. \quad (\text{C6})$$

This general expression holds for an asymmetric jump distribution or, equivalently, for a general reward distribution. To get a typical behavior, we approximate the jump distribution by a Gaussian of mean μ and unit variance. Applying it to Eq. (C6) gives

$$\tau_{\text{fpt}} = \exp \left[\sum_1^\infty \frac{1}{2n} \text{erfc} \left(\frac{\sqrt{n}}{\sqrt{2}} (\hat{\mu}_{M_i} - \mu) \right) \right]. \quad (\text{C7})$$

There is no exact expression for Eq. (C7); we hence focus on extracting the scaling of τ_{fpt} . We first focus on the limit case where $\mu \sim \hat{\mu}_{M_i}$. In this regime, the erfc function will behave as $1/2$ until a cutoff $\sqrt{n_c}(\hat{\mu}_{M_i} - \mu)^{-1} \sim O(1)$. Hence, the leading order behaves as

$$\begin{aligned} \tau_{\text{fpt}} = & \exp \left[\frac{1}{2} \ln \left(\frac{1}{(\hat{\mu}_{M_i} - \mu)^2} \right) + O(1) \right] \\ = & O([\hat{\mu}_{M_i} - \mu]^{-1}). \end{aligned} \quad (\text{C8})$$

This regime includes all cases where the probability of observing diverging τ_{fpt} is the highest. In cases where $\hat{\mu}_{M_i} - \mu = O(1)$, the algorithm will mostly come back to $\hat{\mu}_{M_i}$ after a few steps, which is also given by Eq. (C8). Hence, we assume to extend this approximation in all regimes, leading to the τ_{fpt} scaling given in the main text.

$$\tau_{\text{fpt}}(\mu) = \frac{c}{\hat{\mu}_{M_i} - \mu}. \quad (\text{C9})$$

Combining Eqs. (C1) and (C9) altogether gives the following decision criteria:

$$\begin{aligned} \Delta = & \tilde{R}_{\text{exp}} - \tilde{R}_g \\ = & \tau \int_{\hat{\mu}_{M_i}}^{\mu_{\text{sup}}} \Gamma(\mu)(\mu_* - \mu) d\mu - \tau(\mu_* - \hat{\mu}_{M_i}) \\ & + \int_{\mu_{\text{inf}}}^{\hat{\mu}_{M_i}} \Gamma(\mu)[(\mu_* - \mu)\tau_{\text{fpt}} + (\mu_* - \hat{\mu}_{M_i})(\tau - \tau_{\text{fpt}})] d\mu \\ = & \tau \int_{\hat{\mu}_{M_i}}^{\mu_{\text{sup}}} \Gamma(\mu)(\hat{\mu}_{M_i} - \mu) d\mu \dots \\ & + \int_{\mu_{\text{inf}}}^{\hat{\mu}_{M_i}} \Gamma(\mu)(\hat{\mu}_{M_i} - \mu)\tau_{\text{fpt}} d\mu \\ = & \tau \int_{\hat{\mu}_{M_i}}^{\mu_{\text{sup}}} \Gamma(\mu)(\hat{\mu}_{M_i} - \mu) d\mu + c \int_{\mu_{\text{inf}}}^{\hat{\mu}_{M_i}} \Gamma(\mu) d\mu. \end{aligned} \quad (\text{C10})$$

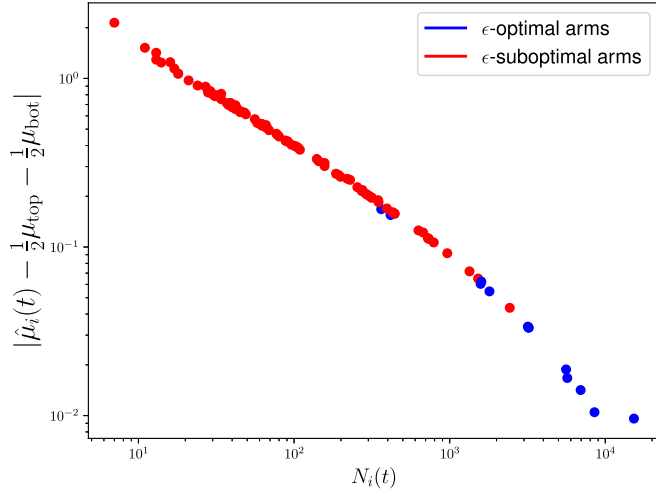


FIG. 7. Distribution of the arms' estimated means according to their respective number of draws. Results are shown for our algorithm at the stopping time with $\epsilon = 0.05$, $m = 10$, $n = 100$, and $\delta = 0.002$. The ϵ -optimal arms, among which m should be returned, are shown in blue, while the remaining arms are displayed in red. On the y axis, the arms' means are subtracted from $(\{\mu_{\mathcal{D}_i}\} + \{\mu_{\mathcal{M}_i}\})/2$, highlighting that the arms closest to the separation line are the ones most frequently drawn.

For a uniform prior, it resumes as the expression given in the main text:

$$\Delta = c \frac{\hat{\mu}_{M_t} - \mu_{\inf}}{\mu_{\sup} - \mu_{\inf}} - \frac{\tau}{2} \frac{(\mu_{\sup} - \hat{\mu}_{M_t})^2}{\mu_{\sup} - \mu_{\inf}}. \quad (\text{C11})$$

Finally, for a uniform prior on $[0, 1]$, it simplifies into

$$\Delta = c \hat{\mu}_{M_t} - \tau \frac{(1 - \hat{\mu}_{M_t})^2}{2}. \quad (\text{C12})$$

Heuristic arguments supporting asymptotic optimality

To further support the empirical evidence of ARM optimality, we consider the asymptotic regime where the algorithms have just converged on a promising arm, given by $\Delta \approx 0$ in Eq. (C12), which is fulfilled when

$$\tau(1 - \hat{\mu}_{M_t})^2 \sim O(1). \quad (\text{C13})$$

We denote by $t_f = T - \tau$ the time spent and by n_f the current number of tested arms. We assume that n_f scales as $O(t_f)$ as long as the promising arm is not found. For a uniform prior, and assuming that the promising arm belongs to the best arms of the subset, the expected regret $(1 - \hat{\mu}_{M_t})$ will thus scale as $1 - \mu_i \sim 1/n_f \sim 1/t_f$. Injecting such scaling into Eq. (C13) leads to $\tau t_f^2 \sim 1$ reached for the first time when $t_f \sim \sqrt{T}$. Hence, the past regret will scale as $t_f \sim \sqrt{T}$ and the resulting upcoming regret of exploiting $\hat{\mu}_{M_t}$ will scale as $\sqrt{\tau} \sim \sqrt{T}$, ensuring the expected optimal scaling. To derive a complete proof, it will be necessary to ensure that the algorithm does not fail to identify $\hat{\mu}_{M_t}$ when tested and no longer spends most of its time trying to identify $\hat{\mu}_{M_t}$ within the arm subset.

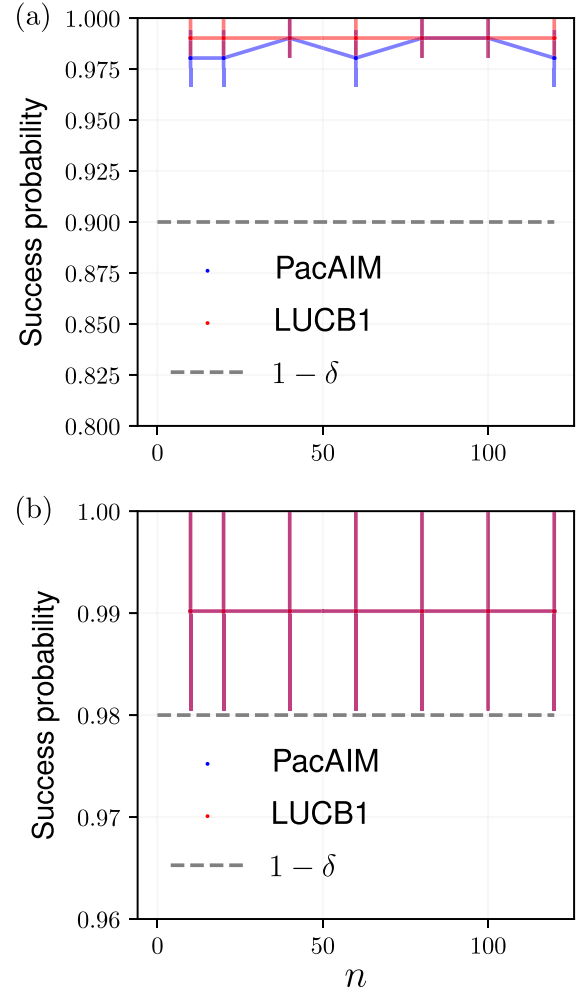


FIG. 8. Success probability of the Explore- m experiments presented in the main text. The probability of identifying the ϵ -optimal subset is averaged over 100 runs for different numbers of arms n . Results are shown for our algorithm (blue) and LUCB1 [41] (red). The ϵ -optimal subset size satisfies $m = n/5$. (a) Mean reward values are sampled from two disjoint ensembles $[0, 0.8]$ and $[0.9, 1]$, with confidence parameters $\epsilon = 0.05$ and $\delta = 0.1$. (b) Suboptimal mean reward values are fixed at 0.7, while ϵ -optimal rewards are set to 0.8, with confidence parameters $\epsilon = 0.05$ and $\delta = 0.02$. Both experiments demonstrate an empirical success probability exceeding the tolerance threshold $1 - \delta$ (gray horizontal lines), validating the stopping time measured in the main text.

APPENDIX D: NUMERICAL EXPERIMENTS

1. Numerical settings

a. Explore- m bandit setting

We here consider Gaussian rewards with $\sigma = 1$. The ϵ -optimal subset size verifies $m = n/5$, with an increasing number of arms tested. In Fig. 3(a), the mean reward values are sampled from two disjoint ensembles $[0, 0.8]$ and $[0.9, 1]$ for the ϵ -optimal subset, with confidence parameters $\epsilon = 0.05$ and $\delta = 0.1$. In Fig. 3(b), suboptimal mean reward values are all set to 0.7, while ϵ -optimal to 0.8, with confidence parameters $\epsilon = 0.05$ and $\delta = 0.02$. For both experiments, the time is averaged over 100 runs with standard deviation indicated.

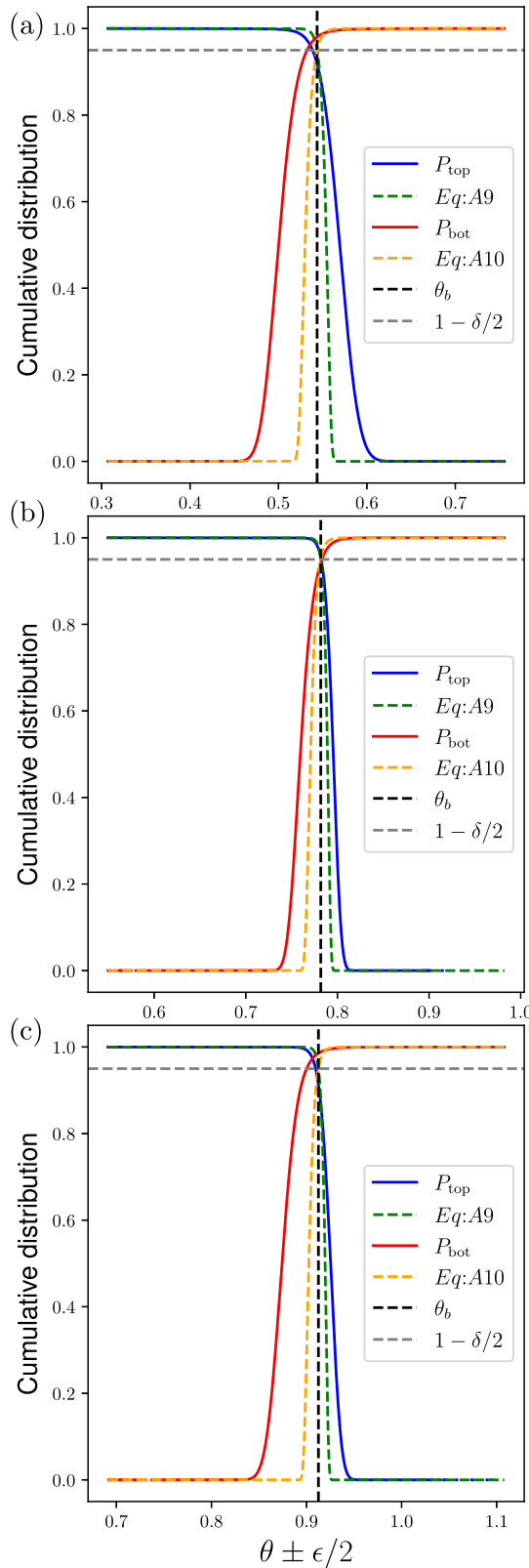


FIG. 9. Cumulative distributions of P_{bot} and P_{top} and their approximations enabling the derivation of θ_b . To illustrate their intersection, the cumulative distributions and their respective approximations are shifted by $\pm\epsilon/2$ ($-\epsilon/2$ for P_{top} and $+\epsilon/2$ for P_{bot}). P_{bot} and P_{top} are shown in blue and red, with their heuristic approximations in green and yellow, respectively. Our approximation returns the parameters $\{\theta_b, t_{\text{top}}, t_{\text{bot}}\}$, such that the approximated Gumbel

Here, seeds are not shared through the algorithms within the same experiment.

b. Linear bandit setting

We here again consider rewards subject to Gaussian noise with $\sigma = 1$. In the first experiment, the arms composing $\mathcal{A}_t$ are sampled from a normal distribution in $\mathbb{R}^{10}$. θ_* is assigned a unit norm. The seeds are taken independently for each algorithm. The regret, averaged over 500 runs, already shows a clear and consistent standard deviation. In the second experiment, $\mathcal{A}_t$ is constituted by two fixed sets presented to the agent with equal probability: the first set $\{[1, 0, 0], [0, 1, 0], [1 - \xi, 2\xi, 0]\}$, and the second set $\{[0, 0.6, 0.8], [0, 0, 1], [0, \xi/10, 1 - \xi]\}$. Finally, $\xi = 0.1$ and $\theta^* = [1, 0, 1]$. The regret, averaged over 50 000 runs, already shows a clear and consistent standard deviation. Here, seeds are shared through the algorithms within the same experiment.

c. Many-armed bandit setting

Here, we consider Bernoulli rewards with a uniform prior. In Fig. 11(a), the parameters used are $T = 10\,000$, $K = 5000 \gg \sqrt{T}$, and $c = 1$. The regret is averaged over 2000 runs. In Fig. 11(b), the mean regret performance when the horizon is reached is given for distinct games with varying horizon length ($K = T \gg \sqrt{T}$ and $c = 1$). The regret is averaged over 2000 runs for each experiment. Here, seeds are shared through the algorithms within the same experiment.

2. Overview of baseline algorithms

We will here review the algorithm used for numerical comparison.

laws should exactly intersect at $(\theta_b, 1 - \delta/2)$, illustrated by the gray and black dashed lines. Results are shown for our algorithm at the stopping time with $\epsilon = 0.05$ and $\delta = 0.1$. In panel (a), $m = 20$, $n = 100$, and the mean rewards are drawn uniformly from $[0, 0.5]$ and $[0.5, 1.0]$ for the suboptimal and optimal sets, respectively; in panel (b), $m = 20$, $n = 200$, and mean rewards are fixed to 0.7 and 0.8, respectively; and in panel (c), $m = 10$, $n = 100$, and mean rewards are drawn uniformly over $[0, 1]$. Note that, in all three cases, the approximate distributions are narrower than the true distributions since the arms have not been sampled uniformly (according to t_{top} and t_{bot}), leading to more spread out distributions. This happens even when the mean rewards are fixed at 0.7 for suboptimal and 0.8 for ϵ -optimal values [panel (b)]. This discrepancy arises because the algorithm still does not sample the means uniformly, as the empirical means observed can differ from the true means in the short term. Hence, although such cases are expected to converge asymptotically (i.e., in the joint limit as $m \rightarrow \infty$, $d \rightarrow \infty$, $\delta \rightarrow 0$), we observe that the stopping time is often reached earlier. When both P_{top} and P_{bot} closely mirror their respective approximations, their actual intersection is near θ_b , as shown in panels (a) and (b). It occurs when both cumulatively account for the same probability of incorrect identification for each set $\mathcal{M}_t$ and $\mathcal{D}_t$. However, since the increments are evaluated for arms individually rather than as a set, P_{top} and P_{bot} may differ at θ_b , depending on the underlying mean distribution, as illustrated in panel (c).

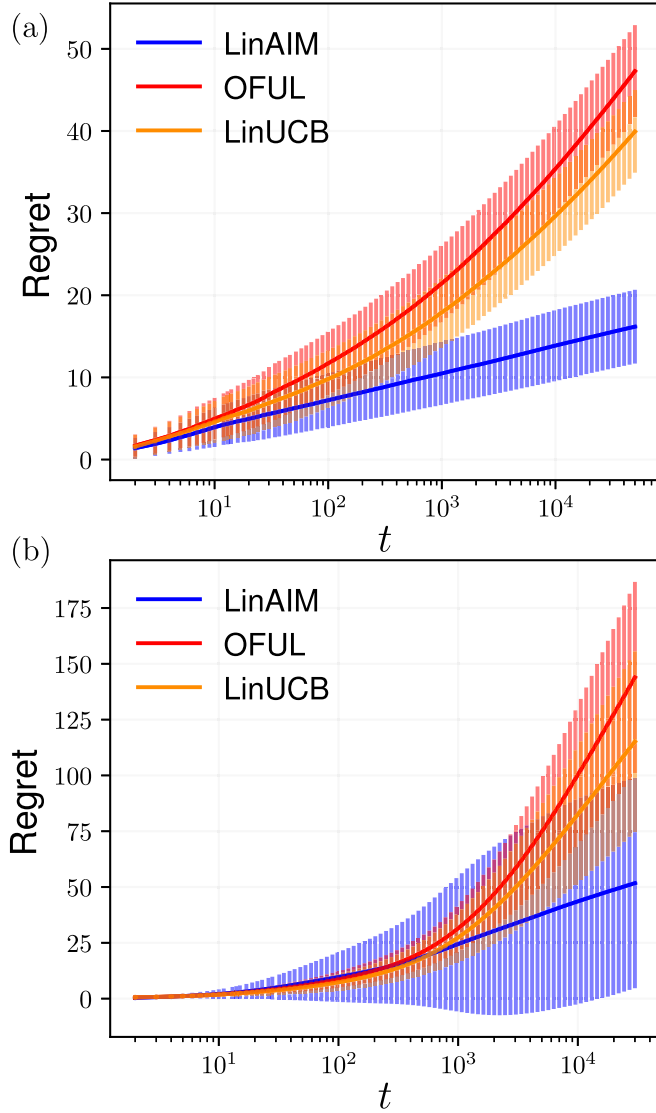


FIG. 10. Mean and standard deviation of the regret for two linear bandit settings with Gaussian rewards. This figure is identical to the one given in the main text but with standard deviation shown with error bars. In blue our algorithm and in orange and red UCB algorithms adjusted to the linear settings. (a) Bandit setting with ten arms resampled at each time from a normal distribution in $\mathbb{R}^{10}$. (b) Toy problem borrowed from Ref. [64] with two contexts. Each context (detailed in the main text) is drawn with equal probability and $\xi = 0.5$. Details of the algorithms, simulations, and a focus on standard deviations are provided in Appendix D.

a. Top-k identification: LUCB1

We rely on the algorithm version first introduced in Ref. [41]. It falls under the category of UCB algorithms, which select the arm maximizing a proxy function typically defined as $F_i = \hat{\mu}_i + C_i(t_i, t)$. In the specific case of subset selection, it reformulates as follows. We define B_i as

$$C_i(t_i, t) = \sqrt{\frac{1}{2t_i} \ln \left[\frac{5nt^4}{4\delta} \right]}. \quad (\text{D1})$$

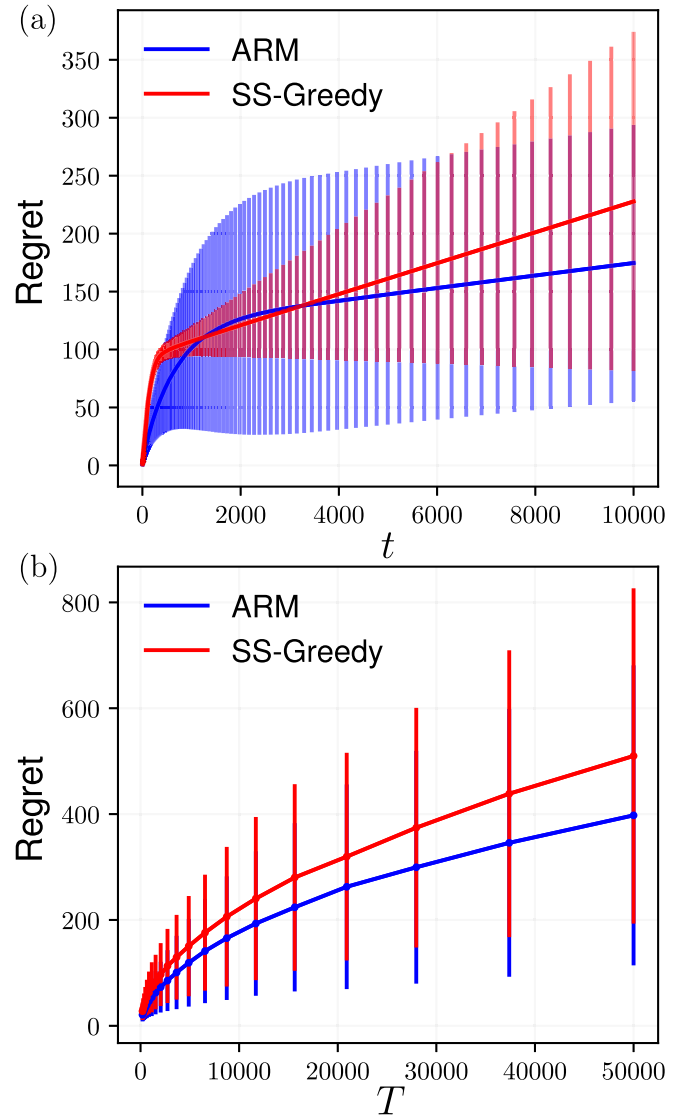


FIG. 11. Mean and standard deviation of the regret for the many-armed bandit problem with Bernoulli rewards and a uniform prior on $[0, 1]$. This figure is identical to the one given in the main text but with standard deviation shown with error bars. (a) The regret growth of our algorithm (ARM), in blue, is observed until horizon T , i.e., the stopping time, is reached $T = 10\,000$ ($K = 5000 \gg \sqrt{T}$ and $c = 1$). The regret is averaged over 2000 runs and standard errors are indicated (see Appendixes D and D2 for details on the numerical settings). Its performance is compared to the ss-greedy algorithm, in red. The final slopes of both algorithms differ, indicating that the selected arm for most of the exploitation is less efficient than the one found for our algorithm. Additionally, our algorithm presents a slower initial slope because exploration and exploitation are already mixed. (b) Mean regret performance when at the stopping time (when the horizon is reached), for distinct games with varying horizon length ($K = T \gg \sqrt{T}$ and $c = 1$). The regret is averaged over 2000 runs.

After having sorted the arms according to their current empirical means and having formed $\mathcal{M}_t$ and $\mathcal{D}_t$, the algorithm selects two arms, denoted by $l_{\text{top},t}$ and $h_{\text{bot},t}$, respectively, in

$\mathcal{M}_t$ and $\mathcal{D}_t$ such that

$$l_{\text{top},t} = \underset{i \in \mathcal{M}_t}{\operatorname{argmin}} (\hat{\mu}_i - C_i(t_i, t)) \quad (\text{D2})$$

and

$$h_{\text{bot},t} = \underset{i \in \mathcal{D}_t}{\operatorname{argmax}} (\hat{\mu}_i + C_i(t_i, t)). \quad (\text{D3})$$

Then, the algorithm stops if

$$\hat{\mu}_{h_{\text{bot},t}} + C_{h_{\text{bot},t}}(t_{h_{\text{bot},t}}, t) - \hat{\mu}_{l_{\text{top},t}} + C_{l_{\text{top},t}}(t_{l_{\text{top},t}}, t) < \epsilon, \quad (\text{D4})$$

meaning that both sets are sufficiently differentiated. If the algorithm does not stop, both arms $l_{\text{top},t}$ and $h_{\text{bot},t}$ are pulled, and the process is repeated.

b. Linear bandit: LinUCB

As for the subset selection, we provide algorithms falling into the UCB category. These algorithms select the arm maximizing a proxy function typically defined as $F_{i,t} = \hat{\mu}_{i,t} + C_{i,t}$. In the linear case, $\hat{\mu}_i$ stands for the empirical mean of drawing arm i at t and C_i an associated confidence ellipsoid. The first LinUCB-type algorithm considered in the numerical experiments, referred to as OFUL [90], is given by

$$F_{i,t} = \hat{\theta}_t' A_t + [\sqrt{d \ln(n(1+t))} + 1] \sqrt{A_t' B_t^{-1} A_t}, \quad (\text{D5})$$

where d stands for the dimension of the arms space, n the arm number at each time step, and B_t and $\hat{\theta}_t$ are defined as in the main text. The second version borrowed from Ref. [26], denoted by LinUCB here, reads

$$F_{i,t} = \hat{\theta}_t' A_t + [\sqrt{2 \ln(n) + \ln(|B_t|)} + 1] \sqrt{A_t' B_t^{-1} A_t}, \quad (\text{D6})$$

where $||$ denotes the determinant operator. We now briefly discuss several potential candidate algorithms that were not implemented in this work. One such algorithm is Lin-Thompson, a refined version of the popular Thompson sampling approach. However, the linear variant of Lin-Thompson, which comes with theoretical guarantees, has a critical drawback: its posterior distribution's variance is adjusted by a factor of $\ln(t)$ [48]. This adjustment leads to overexploration and, consequently, significant regret losses, rendering the algorithm inadequate for meaningful performance comparisons. In addition to Lin-Thompson, we could consider comparing our algorithm with other state-of-the-art

methods such as SOLID [64], OSSB [88], OAM [91], and IDS [95]. Nevertheless, these algorithms suffer from implementation complexities that are not addressed in our current work.

c. Many arms: SS-epsilon greedy

The subsampled greedy algorithm for many-armed settings randomly selects a subset of m arms and performs a greedy strategy after pulling all the arms once. Its pseudocode reads:

- (1) Select m arms uniformly at random without replacement from $[K]$ and pull each arm once.
- (2) Until the end pulls $a_t = \underset{i \in [K]}{\operatorname{argmax}} \hat{\mu}_i$.

The size of the subset is the only parameter that is adjusted depending on the horizon sizes and the reward distribution. Reference [98] proposes $m \sim T^{\beta/(1+\beta)}$ for β -regular prior, i.e., $\mathbb{P}_T[\mu > 1 - \epsilon] = O(\epsilon^\beta)$.

APPENDIX E: ADDITIONAL RESULTS

This Appendix provides additional information and results on our proposed approximate information-maximization algorithms for the three extended multiarmed bandit settings discussed in the main text.

1. Explore- m bandits

Figure 7 focuses on the relation between the number of draws for each arm depending on its empirical mean.

Figure 8 focuses on the success probability of the Explore- m experiments presented in the main text to validate our algorithm.

Figure 9 compares the cumulative posterior distributions P_{bot} and P_{top} to their approximations by Gumbel distributions [Eqs. (A11) and (A12), respectively] and the approximate separator between the two sets, θ_b , for three scenarios where the arms in the optimal and suboptimal sets follow different distributions.

2. Linear bandits

Figure 10 focuses on the standard deviation of the regret in the linear bandit setting.

3. Many-armed bandits

Figure 11 focuses on the standard deviation of the regret in the linear bandit setting.

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction, Adaptive Computation and Machine Learning Series* (A Bradford Book, Cambridge, MA, 1998).
[2] J. I. Gold and M. N. Shadlen, The neural basis of decision making, *Annu. Rev. Neurosci.* **30**, 535 (2007).
[3] M. Jepma and S. Nieuwenhuis, Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory, *J. Cognit. Neurosci.* **23**, 1587 (2011).
[4] H. Kim, J. H. Sul, N. Huh, D. Lee, and M. W. Jung, Role of striatum in updating values of chosen actions, *J. Neurosci.* **29**, 14701 (2009).

[5] M. Vergassola, E. Villermaux, and B. I. Shraiman, ‘Infotaxis’ as a strategy for searching without gradients, *Nature (London)* **445**, 406 (2007).
[6] J. Murlis, J. S. Elkinton, and R. T. Cardé, Odor plumes and how insects use them, *Annu. Rev. Entomol.* **37**, 505 (1992).
[7] J.-B. Masson, M. B. Bechet, and M. Vergassola, Chasing information to search in random environments, *J. Phys. A: Math. Theor.* **42**, 434009 (2009).
[8] Z. Wang and S. Jegelka, Max-value entropy search for efficient bayesian optimization, in *Proceedings of the 34th International Conference on Machine Learning* (Microtome Publishing, Paris, 2017), pp. 3627–3635.

- [9] D. Russo and B. Van Roy, Learning to optimize via information-directed sampling, in *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., Red Hook, New York, 2014), Vol. 27, pp. 1583–1591.
- [10] D. Lindner, M. Turchetta, S. Tschischek, K. Ciosek, and A. Krause, Information directed reward learning for reinforcement learning, *35th Conference on Neural Information Processing Systems (NeurIPS)* (Curran Associates, Inc., Red Hook, New York, 2021).
- [11] R. Houthoofd, X. Chen, X. Chen, Y. Duan, J. Schulman, F. D. Turck, and P. Abbeel, VIME: Variational information maximizing exploration, *Advances in Neural Information Processing Systems (NIPS)*, Vol. 29 (Curran Associates, Inc., Red Hook, New York, 2016), pp. 1109–1117.
- [12] K. Ding, J. Li, and H. Liu, Interactive anomaly detection on attributed networks, in *Proceedings of the 12th ACM International Conference on Web Search and Data Mining, WSDM '19* (Association for Computing Machinery, New York, 2019), pp. 357–365.
- [13] S. Zhang, D. Martinez, and J.-B. Masson, Multi-Robot searching with sparse binary cues and limited space perception, *Front. Rob. AI* **2**, 317 (2015).
- [14] D. Hernandez-Lobato, J. Hernandez-Lobato, A. Shah, and R. Adams, Predictive entropy search for multi-objective bayesian optimization, in *Proceedings of the 33rd International Conference on Machine Learning*, edited by M. F. Balcan and K. Q. Weinberger (PMLR, New York, 2016), Vol. 48, pp. 1492–1501.
- [15] D. Martinez, L. Arhidi, E. Demondion, J. Masson, and P. Lucas, Using insect electroantennogram sensors on autonomous robots for olfactory searches, *J. Vis. Exp.* **90**, e51704 (2014).
- [16] R. T. Cardé, Navigation along windborne plumes of pheromone and resource-linked odors, *Annu. Rev. Entomol.* **66**, 317 (2021).
- [17] J. D. Cohen, S. M. McClure, and A. J. Yu, Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration, *Philos. Trans. R. Soc. B* **362**, 933 (2007).
- [18] K. Doya, *Bayesian Brain: Probabilistic Approaches to Neural Coding* (MIT Press, Cambridge, 2007).
- [19] T. T. Hills, P. M. Todd, D. Lazer, A. D. Redish, and I. D. Couzin, Exploration versus exploitation in space, mind, and society, *Trends Cognit. Sci.* **19**, 46 (2015).
- [20] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O Doherty, and G. Pezzulo, Active inference and learning, *Neurosci. Biobehav. Rev.* **68**, 862 (2016).
- [21] N. Tishby and N. Zaslavsky, Deep learning and the information bottleneck principle, *IEEE Information Theory Workshop (ITW)* (IEEE, Piscataway, New Jersey, 2015).
- [22] M. Helias and D. Dahmen, *Statistical Field Theory for Neural Networks*, Lecture Notes in Physics (Springer International Publishing, Cham, 2020), Vol. 970.
- [23] T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior* (The MIT Press, Cambridge, 2022).
- [24] J.-B. Masson, Olfactory searches with limited space perception, *Proc. Natl. Acad. Sci. USA* **110**, 11261 (2013).
- [25] G. Reddy, V. N. Murthy, and M. Vergassola, Olfactory sensing and navigation in turbulent environments, *Annu. Rev. Condens. Matter Phys.* **13**, 191 (2022).
- [26] T. Lattimore and C. Szepesvári, *Bandit Algorithms* (Cambridge University Press, Cambridge, 2020).
- [27] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Mastering the game of go with deep neural networks and tree search, *Nature (London)* **529**, 484 (2016).
- [28] D. Bouneffouf, I. Rish, and C. Aggarwal, Survey on applications of multi-armed and contextual bandits, *2020 IEEE Congress on Evolutionary Computation (CEC)* (IEEE, Piscataway, New Jersey, 2020), pp. 1–8.
- [29] G. Burtini, J. Loepky, and R. Lawrence, A survey of online experiment design with the stochastic multi-armed bandit, [arXiv:1510.00757](https://arxiv.org/abs/1510.00757).
- [30] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau, Contextual bandits for adapting treatment in a mouse model of *de novo* carcinogenesis, in *Proceedings of the 3rd Machine Learning for Healthcare Conference* (Microtome Publishing, Paris, 2018), pp. 67–82.
- [31] C. Gentile, S. Li, and G. Zappella, Online clustering of bandits, in *Proceedings of the 31st International Conference on Machine Learning* (Microtome Publishing, Paris, 2014), pp. 757–765.
- [32] V. Kuleshov and D. Precup, Algorithms for multi-armed bandit problems, [arXiv:1402.6028](https://arxiv.org/abs/1402.6028).
- [33] J. Mary, R. Gaudel, and P. Preux, Bandits and recommender systems, in *Machine Learning, Optimization, and Big Data*, Lecture Notes in Computer Science, edited by P. Pardalos, M. Pavone, G. M. Farinella, and V. Cutello (Springer International Publishing, Cham, 2015), pp. 325–336.
- [34] T. C. Stewart, T. Bekolay, and C. Eliasmith, Learning to select actions with spiking neurons in the basal ganglia, *Front. Neurosci.* **6**, 1 (2012).
- [35] T. L. Lai and H. Robbins, Asymptotically efficient adaptive allocation rules, *Adv. Appl. Math.* **6**, 4 (1985).
- [36] E. Kaufmann, N. Korda, and R. Munos, Thompson sampling: An asymptotically optimal finite-time analysis, in *Algorithmic Learning Theory*, Lecture Notes in Computer Science, edited by N. H. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann (Springer, Berlin, 2012), pp. 199–213.
- [37] P. Auer, Using upper confidence bounds for online learning, in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, Redondo Beach, CA, 2000), pp. 270–279.
- [38] S. Bubeck, R. Munos, and G. Stoltz, Pure exploration in finitely-armed and continuous-armed bandits, *Theor. Comput. Sci.* **412**, 1832 (2011).
- [39] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, LIL' UCB: An optimal exploration algorithm for multi-armed bandits, in *Proceedings of the 27th Conference on Learning Theory* (Microtome Publishing, Paris, 2014), pp. 423–439.
- [40] E. Kaufmann and S. Kalyanakrishnan, Information complexity in bandit subset selection, in *Proceedings of the 26th Annual Conference on Learning Theory* (Microtome Publishing, Paris, 2013), pp. 228–251.

- [41] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone, PAC subset selection in stochastic multi-armed bandits, in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* (Microtome Publishing, Paris, 2012).
- [42] R. Degenne and W. M. Koolen, Pure exploration with multiple correct answers, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Red Hook, New York, 2019), Vol. 32, pp. 14591–14600.
- [43] L. Li, W. Chu, J. Langford, and R. E. Schapire, A contextual-bandit approach to personalized news article recommendation, in *Proceedings of the 19th International Conference on World Wide Web, WWW '10* (Association for Computing Machinery, New York, 2010), pp. 661–670.
- [44] B. Lin and D. Bouneffouf, Optimal epidemic control as a contextual combinatorial bandit with budget, *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (IEEE, Piscataway, New Jersey, 2022), pp. 1–8.
- [45] L. Zhou, A survey on contextual multi-armed bandits, [arXiv:1508.03326](https://arxiv.org/abs/1508.03326).
- [46] P. Auer, Using confidence bounds for exploitation-exploration trade-offs, *J. Mach. Learn. Res.* **3**, 397 (2002).
- [47] T. Lattimore and C. Szepesvari, The end of optimism? An asymptotic analysis of finite-armed linear bandits, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Microtome Publishing, Paris, 2017), pp. 728–737.
- [48] S. Agrawal and N. Goyal, Thompson sampling for contextual bandits with linear payoffs, in *Proceedings of the 30th International Conference on Machine Learning* (Microtome Publishing, Paris, 2013), pp. 127–135.
- [49] S. S. Villar, J. Bowden, and J. Wason, Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges, *Stat. Sci.* **30**, 199 (2015).
- [50] S. S. Villar, Bandits strategies evaluated in the context of clinical trials in rare life-threatening diseases, *Prob. Eng. Inf. Sci.* **32**, 229 (2018).
- [51] R. de Heide, J. Cheshire, P. Ménard, and A. Carpentier, Bandits with many optimal arms, in *Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21* (Curran Associates Inc., Red Hook, NY, 2024), pp. 22457–22469.
- [52] T. Bonald and A. Proutiere, Two-target algorithms for infinite-armed bandits with bernoulli rewards, in *Advances in Neural Information Processing Systems*, edited by C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., Red Hook, New York, 2013), Vol. 26, pp. 2184–2192.
- [53] D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, and L. A. Shepp, Bandit problems with infinitely many arms, *Ann. Stat.* **25**, 2103 (1997).
- [54] Y. Wang, J.-y. Audibert, and R. Munos, Algorithms for infinitely many-armed bandits, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, New York, 2008), Vol. 21, pp. 1729–1736.
- [55] D. Baudry, E. Kaufmann, and O.-A. Maillard, Sub-sampling for efficient non-parametric bandit exploration, in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20* (Curran Associates Inc., Red Hook, NY, 2020), pp. 5468–5478.
- [56] S. Guha, K. Munagala, and P. Shi, Approximation algorithms for restless bandit problems, *J. ACM* **58**, 1 (2010).
- [57] N. Levine, K. Crammer, and S. Mannor, Rotting bandits, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2017), Vol. 30, pp. 3074–3083.
- [58] P. Auer, N. Cesa-Bianchi, and P. Fischer, Finite-time analysis of the multiarmed bandit problem, *Mach. Learn.* **47**, 235 (2002).
- [59] H. Bastani, M. Bayati, and K. Khosravi, Mostly exploration-free algorithms for contextual bandits, *Manage. Sci.* **67**, 1329 (2021).
- [60] P. Ménard and A. Garivier, A minimax and asymptotically optimal algorithm for stochastic bandits, in *Proceedings of the 28th International Conference on Algorithmic Learning Theory* (Microtome Publishing, Paris, 2017), pp. 223–237.
- [61] D. Russo and B. V. Roy, An information-theoretic analysis of thompson sampling, *J. Mach. Learn. Res.* **17**, 1 (2016).
- [62] G. Reddy, A. Celani, and M. Vergassola, Infomax strategies for an optimal balance between exploration and exploitation, *J. Stat. Phys.* **163**, 1454 (2016).
- [63] Y.-H. Hung, P.-C. Hsieh, X. Liu, and P. R. Kumar, Reward-biased maximum likelihood estimation for linear stochastic bandits, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35 (AAAI Press, 2021), pp. 7874–7882.
- [64] A. Tirinzoni, M. Pirotta, M. Restelli, and A. Lazaric, An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits, in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20* (Curran Associates Inc., Red Hook, NY, 2020), pp. 1417–1427.
- [65] A. Barbier-Chebbah, C. L. Vestergaard, and J.-B. Masson, Approximate information for efficient exploration-exploitation strategies, *Phys. Rev. E* **109**, L052105 (2024).
- [66] A. B. Barbier Chebbah, C. L. Vestergaard, J.-B. Masson, and E. Boursier, Approximate information maximization for bandit games, in *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Microtome Publishing, Paris, 2025), pp. 316–324.
- [67] A. Slivkins, Introduction to multi-armed bandits, *Found. Trends Mach. Learn.* **12**, 1 (2019).
- [68] E. Kaufmann, O. Cappé, and A. Garivier, On the complexity of best-arm identification in multi-armed bandit models, *J. Mach. Learn. Res.* **17**, 1 (2016).
- [69] S. Kalyanakrishnan and P. Stone, Efficient selection of multiple bandit arms: Theory and practice, in *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10* (Omnipress, Madison, WI, 2010), pp. 511–518.
- [70] S. Mannor and J. N. Tsitsiklis, The sample complexity of exploration in the multi-armed bandit problem, *J. Mach. Learn. Res.* **5**, 623 (2004).
- [71] Y. Zhou, X. Chen, and J. Li, Optimal PAC multiple arm identification with applications to crowdsourcing, in *Proceedings of the 31st International Conference on Machine Learning* (Microtome Publishing, Paris, 2014), pp. 217–225.

- [72] S. Bubeck, T. Wang, and N. Viswanathan, Multiple identifications in multi-armed bandits, in *Proceedings of the 30th International Conference on Machine Learning* (Microtome Publishing, Paris, 2013), pp. 258–265.
- [73] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, Best arm identification: A unified approach to fixed budget and fixed confidence, in *Advances in Neural Information Processing Systems* edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., Red Hook, New York, 2012), Vol. 25, pp. 3212–3220.
- [74] J.-Y. Audibert and S. Bubeck, Best arm identification in multi-armed bandits, in *COLT—23th Conference on Learning Theory* (Omnipress, Madison, Wisconsin, 2010), p. 13.
- [75] Note that ϵ is fixed and thus independent of the arms' history.
- [76] S. N. Majumdar, A. Pal, and G. Schehr, Extreme value statistics of correlated random variables: A pedagogical review, *Phys. Rep.* **840**, 1 (2020).
- [77] A. Carpentier and M. Valko, Simple regret for infinitely many armed bandits, in *Proceedings of the 32nd International Conference on Machine Learning* (Microtome Publishing, Paris, 2015), pp. 1133–1141.
- [78] I. Rejwan and Y. Mansour, Top- k combinatorial bandits with full-bandit feedback, in *Proceedings of the 31st International Conference on Algorithmic Learning Theory* (Microtome Publishing, Paris, 2020), pp. 752–776.
- [79] S. Katariya, A. Tripathy, and R. Nowak, MaxGap bandit: Adaptive algorithms for approximate ranking, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2019), pp. 11047–11057.
- [80] A. Locatelli, M. Gutzeit, and A. Carpentier, An optimal algorithm for the thresholding bandit problem, in *Proceedings of the 33rd International Conference on Machine Learning* (Microtome Publishing, Paris, 2016), pp. 1690–1698.
- [81] B. Mason, L. Jain, A. Tripathy, and R. Nowak, Finding all ϵ -good arms in stochastic bandits, in *Advances in Neural Information Processing Systems* edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, NY, 2020), Vol. 33, pp. 20707–20718.
- [82] W. Ren, J. Liu, and N. B. Shroff, Exploring k out of top ρ fraction of arms in stochastic bandits, in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* (Microtome Publishing, Paris, 2019), pp. 2820–2828.
- [83] H. Zhuang, C. Wang, and Y. Wang, Identifying outlier arms in multi-armed bandit, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, New York, 2017), Vol. 30, pp. 5204–5213.
- [84] N. Abe, A. W. Biermann, and P. M. Long, Reinforcement learning with immediate rewards and linear hypotheses, *Algorithmica* **37**, 263 (2003).
- [85] W. Chu, L. Li, L. Reyzin, and R. Schapire, Contextual bandits with linear payoff functions, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Microtome Publishing, Paris, 2011), pp. 208–214.
- [86] A. Slivkins, Contextual bandits with similarity information, in *Proceedings of the 24th Annual Conference on Learning Theory* (Microtome Publishing, Paris, 2011), pp. 679–702.
- [87] The definition of pseudoregret has not yet reached a consensus in the community. However, from a numerical perspective, all definitions agree on the objective of minimizing the average pseudoregret.
- [88] R. Combes, S. Magureanu, and A. Proutiere, Minimal exploration in structured stochastic bandits, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Curran Associates Inc., Red Hook, NY, 2017), pp. 1761–1769.
- [89] V. Dani, T. P. Hayes, and S. M. Kakade, Stochastic linear optimization under bandit feedback, in *21st Annual Conference on Learning Theory—COLT 2008, Helsinki, Finland, 9–12 July 2008*, edited by R. A. Servedio and T. Zhang (Omnipress, Madison, Wisconsin, 2008), pp. 355–366.
- [90] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári, Improved algorithms for linear stochastic bandits, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, New York, 2011), Vol. 24, pp. 2312–2320.
- [91] B. Hao, T. Lattimore, and C. Szepesvari, Adaptive exploration in linear contextual bandit, *Proceedings of the Twenty Third-International Conference on Artificial Intelligence and Statistics*, Vol. 108 (PMLR, 2020), pp. 3536–3545.
- [92] P. Rusmevichientong and J. N. Tsitsiklis, Linearly parameterized bandits, *Math. Oper. Res.* **35**, 395 (2010).
- [93] Note that these results rely on an adjustment of the variance of the posterior distribution.
- [94] M. Abeille and A. Lazaric, Linear thompson sampling revisited, *Electron. J. Stat.* **11**, 5165 (2017).
- [95] J. Kirschner, T. Lattimore, C. Vernade, and C. Szepesvari, Asymptotically optimal information-directed sampling, in *Proceedings of 34th Conference on Learning Theory* (Microtome Publishing, Paris, 2021), pp. 2777–2821.
- [96] J. Kirschner and A. Krause, Information directed sampling and bandits with heteroscedastic noise, in *Proceedings of the 31st Conference on Learning Theory* (Microtome Publishing, Paris, 2018), pp. 358–384.
- [97] J. Kirschner, T. Lattimore, and A. Krause, Information directed sampling for linear partial monitoring, in *Proceedings of 33rd Conference on Learning Theory* (Microtome Publishing, Paris, 2020), pp. 2328–2369.
- [98] M. Bayati, N. Hamidi, R. Johari, and K. Khosravi, Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms, in *Advances in Neural Information Processing Systems* edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., Red Hook, NY, 2020), Vol. 33, pp. 1713–1723.
- [99] K. Jamieson and L. Jain, A bandit approach to multiple testing with false discovery control, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18* (Curran Associates Inc., Red Hook, NY, 2018), pp. 3664–3674.
- [100] The lower bounds hold for sufficiently high T, K values and 1-regular Γ ($\mu_{\sup} = 1$), meaning that $\mathcal{P}_\Gamma[\mu > 1 - \epsilon] = O(\epsilon)$ as ϵ goes to 0; see Ref. [98] for theorem details.

- [101] Note that a sequential strategy has been proposed in Ref. [98], but this strategy still depends on a fixed bound on the subset size.
- [102] <https://gitlab.Pasteur.Fr/decision-making-related/papers/BroadBanditGame>.
- [103] E. S. Andersen, On the fluctuations of sums of random variables II, *Math. Scand.* **2**, 195 (1954).
- [104] M. Bauer, C. Godrèche, and J. M. Luck, Statistics of persistent events in the binomial random walk: Will the drunken sailor hit the sober man? *J. Stat. Phys.* **96**, 963 (1999).
- [105] P. Le Doussal and K. J. Wiese, Driven particle in a random landscape: Disorder correlator, avalanche distribution, and extreme value statistics of records, *Phys. Rev. E* **79**, 051105 (2009).