Bandits with many optimal arms

Rianne de Heide
INRIA Lille and CWI Amsterdam
r.de.heide@cwi.nl

James Cheshire
Otto von Guericke University Magdeburg
james.cheshire@ovgu.de

Pierre Ménard
Otto von Guericke University Magdeburg
pierre.menard@ovgu.de

Alexandra Carpentier
University of Potsdam
carpentier@uni-potsdam.de

Abstract

We consider a stochastic bandit problem with a possibly infinite number of arms. We write $p^\star$ for the proportion of optimal arms and Δ for the minimal mean-gap between optimal and sub-optimal arms. We characterize the optimal learning rates both in the cumulative regret setting, and in the best-arm identification setting in terms of the problem parameters T (the budget), $p^\star$ and Δ. For the objective of minimizing the cumulative regret, we provide a lower bound of order $\Omega(\log(T)/(p^\star\Delta))$ and a UCB-style algorithm with matching upper bound up to a factor of $\log(1/\Delta)$. Our algorithm needs $p^\star$ to calibrate its parameters, and we prove that this knowledge is necessary, since adapting to $p^\star$ in this setting is impossible. For best-arm identification we also provide a lower bound of order $\Omega(\exp(-cT\Delta^2 p^\star))$ on the probability of outputting a sub-optimal arm where $c > 0$ is an absolute constant. We also provide an elimination algorithm with an upper bound matching the lower bound up to a factor of order $\log(T)$ in the exponential, and that does not need $p^\star$ or Δ as parameter. Our results apply directly to the three related problems of competing against the j-th best arm, identifying an ε good arm, and finding an arm with mean larger than a quantile of a known order.

1 Introduction

In the classical stochastic multi-armed bandit model – see [37] for a recent survey – a learner interacts with an environment in several rounds. At each round, the learner chooses an *arm* to play, and receives a random reward from the associated probability distribution. Popular settings are respectively the fixed budget *cumulative regret setting* [41], and *best-arm identification setting* [18, 7, 1]. In the first setting, the learner is interested in maximizing the sum of rewards gathered – or minimizing the cumulative regret – and in the best-arm identification setting, the learner is asked at the end of the game to output a guess for the arm with the largest mean reward, and is interested in the quality of this guess – typically measured by the probability of error in the guess.

In most of the papers that concern this topic, it is assumed (i) that there is a single optimal arm, i.e. arm with highest mean, and (ii) that the number of arms is bounded and small when compared to the time horizon, i.e. the number of rounds where the player is allowed to choose an arm. However in many realistic applications, it is not the case, for example in image classification, mining of resources, personalized medicine, or hyperparameter tuning (see [5] for more examples). And while it is clear that in all generality, the task of the learner becomes unsolvable if the number of arms is too large, it intuitively makes sense that if the proportion of optimal arms is also large, this should help the learner.

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

In this paper, we lift both assumptions summarised in (i) and (ii) and study both the cumulative regret and best-arm identification setting. See Section 1.3 for literature related to this that we will discuss later. We will focus on the *problem dependent setting* and will aim at characterising optimal learning rates depending on the proportion of optimal arms, and on the minimal gap between the mean of an optimal arm and the mean of a sub-optimal arm.

1.1 Setting

We consider a setting with a (potentially infinite) set of arms $\mathcal{A}$, which we call the *reservoir*. Each arm $a \in \mathcal{A}$ is associated with a probability distribution ν_a, which we assume to be supported on $[0,1]$, and we denote its mean by μ_a. Write $\mu^* = \max_{a\in\mathcal{A}} \mu_a$ for the highest mean[1], $\mu_{sub} = \sup_{a\in\mathcal{A}:\mu_a\neq\mu^*} \mu_a$ for the second highest mean, and $\Delta = \mu^* - \mu_{sub}$ for the associated minimal gap. We will focus throughout this paper on the case where $\Delta > 0$.

We further assume that there exists a partition $\mathcal{A} = \mathcal{A}^* \cup \mathcal{A}_{sub}$ such that each arm $a \in \mathcal{A}^*$ is optimal, i.e. $\mu_a = \mu^*$, and each arm $a \in \mathcal{A}_{sub}$ is sub-optimal, i.e. $\mu_a \leq \mu_{sub}$. We assume that the agent can pick arms uniformly at random from the reservoir $\mathcal{A}$[2], and this arm belongs either to the set $\mathcal{A}^*$ with probability $p^\star$, i.e. there is a proportion $p^\star$ of optimal arms in the reservoir; or it belongs to the set $\mathcal{A}_{sub}$ with probability $1-p^\star$, i.e. there is a proportion $1-p^\star$ of sub-optimal arms in the reservoir.

The learner interacts with the environment in several rounds $t = 1, 2, \ldots, T$, where we fix the time horizon T. At each round $t \leq T$, the learner chooses an arm a_t by either picking a new arm from the reservoir $\mathcal{A}$ or playing a past arm, and gets a reward $Y_t \sim \nu_{a(t)}$. The arm choice depends only on the past observations, the past arm choices, and possibly some exogenous randomness. The rewards for each arm a are i.i.d. random variables with mean μ_a unknown to the learner.

Cumulative regret setting. The first setting we study is that of minimizing the *cumulative regret*. This setting enforces the *exploration-exploitation trade-off*: the learner needs to balance exploratory actions to get a better estimate of the reward distributions, and exploitative actions to maximize the total return – and minimise the associated cumulative regret. The cumulative regret is the difference between the sum of expected rewards the learner would have obtained by only choosing the arm with the highest mean reward, and the sum of expected rewards she actually collected:

$$R(T) = \sum_{t=1}^{T} \mu^\star - \mu_{a(t)} \,.$$

Best-arm identification setting In the second setting we study, we are interested in identifying an arm with the highest mean reward. At the end of T rounds, the agents selects an arm $\hat{a}_T$ and aims at minimising the probability of outputting an arm with sub-optimal mean:

$$\mathrm{e}(T) = \mathbb{P}(\hat{a}_T \notin \mathcal{A}^*).$$

A closely related popular measure of error is the *simple regret*, which is not discussed in this paper.

Equivalent settings Firstly, our setting is directly applicable to the problem of competing against the j-th best arm, where we assume w.l.o.g. the arms to be ordered according to their means. Indeed our setting translates to this if we replace $p^\star$ by j/K and Δ by the gap between the $j/2$-th and the $j+1$-th best arm, i.e. $\Delta = |\mu_{j/2} - \mu_{j+1}|$. Secondly, our setting is directly applicable to that of identifying an ε good arm, and thirdly, our setting is directly applicable to finding any arm in the reservoir with a mean larger than the quantile of a known order – see the discussion in Section 1.3.

1.2 Contributions

We characterise the optimal learning rates both for the cumulative regret setting, and for best-arm identification, for our problem described above. We characterise the optimal learning rates in terms of the problem parameters $T, p^\star$, and Δ.

In order to describe our results, let us write for $\bar{\Delta} > 0$, $\bar{p^\star} \in [0,1)$: $\mathfrak{B}_{\bar{\Delta},\bar{p^\star}}$, for the set of bandit problems whose reservoir distribution is such that $p^\star \geq \bar{p^\star}$ and such that $|\bar{\mu}^\star - \mu_{sub}| \geq \bar{\Delta}$.

[1] We assume that it is attained for some arm(s).

[2] In case of infinite $\mathcal{A}$, one can obviously not sample from a uniform distribution. Our analysis extends to general distributions on $\mathcal{A}$.

Cumulative regret We provide an algorithm, *that takes $p^\star$ as a parameter*, that is such that (see Theorem 1)

$$\mathbb{E}R(T) \le O\left(\frac{\log T \log(1/\Delta)}{p^\star \Delta}\right).$$

Conversely, we prove in Theorem 2 that for $\bar{p^\star} \le 1/4$ and $\bar{\Delta} \le 1/4$, and for any algorithm, there exists a problem in $\mathfrak{B}_{\bar{\Delta},\bar{p^\star}}$ such that

$$\mathbb{E}R(T) \ge \Omega\left(\frac{\log T}{\bar{p^\star}\bar{\Delta}}\right).$$

These two bounds match up to a multiplicative factor of order $\log(1/\Delta)$. They highlight the intuitive fact that we should pay the number of arms in the rate only relative to the number of optimal arms – i.e. only through $p^\star$. Indeed, the probability of picking an optimal arm in the reservoir when sampling uniformly at random being $p^\star$, if we sample about $1/p^\star$ arms at random from the reservoir, we will have sampled one optimal arm with constant probability – so that $1/p^\star$ plays the same role as the number of arms.

Having said that, there is a main conceptual difficulty in order to get a rate that is tight in terms of its dependence in T. If we sample only $1/p^\star$ arms from the reservoir, the probability of having no optimal arms in the chosen set of arms is also a constant – so that the regret is linear in T. It is therefore essential to sample *more* arms. In order to have a logarithmic regret in T, we need to sample at least about $\log T/p^\star$ arms from the reservoir – in which case at least one of them is optimal with probability polynomially decaying with T. But if we do this, we get a regret of order $\frac{(\log T)^2}{p^\star\Delta}$, as there are about $\log T/p^\star$ sub-optimal arms whenever $p^\star$ is not too close to 1. This is much larger than the bound that we have, where the dependence on T is only $\log T$. In order to achieve this bound, we need to take into account the fact that when sampling $\log T/p^\star$ arms from the reservoir, there is typically not just 1, but $\log T$ optimal arms with high probability – and leverage this fact both in our algorithm and in the associated proof. We describe this in more detail in Section 2.1.

Best-arm identification We provide an algorithm *that does not take $p^\star$ as a parameter*, such that,

$$\mathrm{e}(T) \le O\left(\log(T)\exp\left(-c\frac{T\Delta^2 p^\star}{\log(T)}\right)\right),$$

where c is some universal constant. Conversely, we prove that for $p^\star \le 1/4$ and $\Delta \le 1/4$, and for any algorithm, there exists a problem in $\mathfrak{B}_{\bar{\Delta},\bar{p^\star}}$ such that $\mathrm{e}(T) \ge \Omega\big(\exp\big(-cT\Delta^2 p^\star\big)\big)$, where $c > 0$ is some universal constant. These two bounds match in order up to a factor of order $\log(T)$ in the exponential, it is an open question here whether this term is necessary or not.

These bounds highlight the intuitive fact that we should pay the number of arms in the rate only relative to the number of optimal arms – i.e. only through $p^\star$. As in the cumulative regret setting, if we sample about $1/p^\star$ arms at random from the reservoir, we will have sampled one optimal arm with constant probability – so that $1/p^\star$ plays the same role as the number of arms.

As in the cumulative regret setting, there is again a main conceptual difficulty in order to get a rate that is tight in terms of its dependence in T. If we sample only $1/p^\star$ arms from the reservoir, the probability of having no optimal arms in the chosen arms is also a constant – which is way smaller than the targeted best-arm identification probability. In order to have at least one optimal arm in the set of arms picked from the reservoir with a probability that decays exponentially with $p^\star T\Delta^2$, the number of arms that have to be sampled should be larger than $T\Delta^2$. But if we do this, we get an upper bound on the probability of error that is of constant order – which is much larger than the bound that we have. In order to obtain our upper bound, we need to take into account the fact that when sampling $T\Delta^2$ arms from the reservoir, there is typically not just 1, but $p^\star T\Delta^2$ optimal arms with high probability – and leverage this fact both in our algorithm and in the associated proof. We describe this in more detail in Section 3.1.

Adaptation to $p^\star$: diverging pictures for cumulative regret and best-arm identification The algorithm for cumulative regret takes (a lower bound on) $p^\star$ as parameter, but the algorithm for best-arm identification does not take anything related to $p^\star$ or Δ as a parameter. And so, while our algorithm for best-arm identification is adaptive to $p^\star$ and Δ, our cumulative regret algorithm is adaptive to Δ but not $p^\star$. In Section 2.3 we prove that it is not just a weakness of our analysis, but that it is *impossible to adapt to $p^\star$ when it comes to the cumulative regret*. The phenomenon of adaptation to the problem hyper-parameters being possible for best-arm identification but not for cumulative regret, was observed earlier: In the $\mathcal{X}$-armed bandit setting [40] show it is impossible to adapt to smoothness and [24] further classifies the cost of adaptation in this case. [44] explore the cost of adaptation to $p^\star$ for the problem independent case where the number of arms is large.

1.3 Related work

Finite and small number of arms. The regret-minimization setting, introduced by [41], has been well-studied for *finite*-armed bandit models. Algorithms for this problem fall into several categories: algorithms based on upper-confidence bounds (UCB) for the unknown arm means [30, 3, 2, 11], algorithms that exploit a posterior distribution on the means, such as Thompson Sampling [42, 34], and many more such as explore-then-commit [20] and phased-elimination [19]. Logarithmic instance-dependent lower bounds have already been obtained in the seminal paper by [36], and were generalized later, e.g. by [10], see [21] for an overview and simple proofs. In the setting where the number of arms $|\mathcal{A}|$ is finite and not too large – much smaller than T – a classical problem dependent upper bound on the expected cumulative regret is[3]

$$\sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \left(\frac{8 \log T}{\mu^* - \mu_a} + 2 \right) \leq |\mathcal{A}_{sub}| \frac{\log T}{\Delta} + 2|\mathcal{A}_{sub}|. \tag{1}$$

The bound in the RHS is tight if all sub-optimal arms have the same gap Δ. Moreover, this regret bound asymptotically matches the lower bound by [10] up to a multiplicative constant. In the case where there are infinitely many sub-optimal arms, on the other hand, this upper bound is infinite, *even when the proportion of optimal arms $p^\star$ is large and where one would hope for better performances*.

The fixed-budget best-arm identification setting was introduced by [7, 1] and has been widely studied. It is well-known that algorithms that are optimal for cumulative-regret minimization cannot yield optimal performance for best-arm identification [8, 33]. Write[3] $H = \sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \frac{1}{(\mu^* - \mu_a)^2} \leq \frac{|\mathcal{A}_{sub}|}{\Delta^2}$. The bound in the RHS is tight if all sub-optimal arms have gap Δ. It is proven by [1] that given H, there exists an algorithm such that the probability of misidentifying an optimal arm is of order $\exp(-cT/H)$, where $c > 0$ is some universal constant. In the case where there is *a single optimal arm* this bound is provably optimal [12] when H is known. However, in the case where there are infinitely many sub-optimal arms this upper bound is larger than 1 and thus vacuous, *even when the proportion of optimal arms $p^\star$ is large and where one would hope for better performances*.

Importantly, our results in both settings extend to finite bandits. Furthermore we do not need infinite $\mathcal{A}$ for our results to be near optimal. In the finite setting with K arms and $p^\star K$ optimal arms the problem is strictly harder than one with $\frac{1}{p^\star}$ arms and a single optimal arm. Indeed, the latter problem would correspond to one where the learner receives, as additional information, a partition of the set of K arms in $\frac{1}{p^\star}$ groups, where one of the groups contains all optimal arms, and the others are only composed of sub-optimal arms. One can then see that we match the classical UB and LB for the finite bandit problem, up to $\log(1/\Delta)$ terms.

Large to infinite number of arms. The setting with an infinite number of arms – and sometimes also many optimal arms – has been studied in different settings.

A setting that is very related to ours is the infinitely many-armed setting where a distribution is assumed on the reservoir – called the reservoir distribution. At each round, the learner can pull a previously queried arm, or a new arm that is sampled according to the reservoir distribution. A classical assumption on the reservoir is that the proportion of $\bar{\Delta}$-near optimal arms is of larger order than $\bar{\Delta}^{-\alpha}$ for any $\bar{\Delta}$. This setting been studied for both cumulative regret minimization [5, 43, 6, 17] and for best-arm identification [13, 4, 15]. A classical strategy is to select a subset of arms from the reservoir, large enough so that it contains a near optimal arm with high probability, and to use classical bandit strategies on these arms. The minimax order of magnitude of the cumulative regret is then $\sqrt{T} \vee T^{\alpha/(\alpha+1)}$ and for the simple regret it is $T^{-1/2} \vee T^{-1/\alpha}$.

[3] In the case where $\mathcal{A}$ is finite otherwise the quantity below is infinite.

Related results have also be obtained in the setting where the number of arms is finite, but large – i.e. $K > T$ – and under related assumptions on the frequency of near-optimal arms [44]. While our setting is extremely related to this setting, the assumption about the frequency of near-optimal arms differs in the above literature from the assumption we make in this paper. Their bounds are not dependent upon Δ – they assume $\forall k \in [K], \mu_k \in [0,1]$, and instead focus on achieving semi adaptivity in regards to an unknown α^*, where $\alpha^* := \inf\{\alpha : K/|S_*| < T^\alpha\}$. In the context of our setting T^α would act as a upper bound on $1/p^\star$. They propose an algorithm with user defined parameter β that has no guarantees on regret for $\beta < \alpha$. And while our assumption is more restrictive, we also expect to obtain much smaller optimal rates. Our results differ from this stream of literature in the same way that, in the classical MAB, *problem dependent results differ from problem independent results*.

Another setting takes a regularity assumption on the reservoir distribution around μ^* – that is, the proportion of arms in the reservoir whose gap is of order greater than $\bar{\Delta}$ is bounded above by a function of $\bar{\Delta}$, typically $\bar{\Delta}^\alpha$, where α is the regularity coefficient. For best-arm identification adaptivity is possible without knowledge of α and [13] provide algorithms for the simple regret with LB matching up to $\log(T)$ terms. In the case of cumulative regret [43] and [6] again provide near optimal results but in the case of *known* α. While the above literature considers a weaker assumption on the reservoir distribution, their results are also considerably weaker than our own. For best-arm identification they identify a sub optimal arm whose distance to the optimal arm is bounded polynomially with T. For cumulative regret the regret is bounded polynomially with T. These bounds are in both cases much larger than our bounds – which essentially reflects that their assumption are weaker.

Closer to our setting are the works [15] and [4], where they try to find any arm in the reservoir with a mean larger than the quantile of a known order (with respect to the reservoir distribution) with high probability. This can be seen as the fixed confidence version of our setting for best-arm identification where the order of the quantiles is our known proportion of optimal arms $p^\star$ and the gap Δ is the difference between the first and the second quantile of order $p^\star$. Precisely, [4] provide an algorithm that can find an arm above the quantile of order $p^\star$ with probability at least $1-\delta$ in less than $H_{\Delta,p^\star}\log(1/\delta)^2$ samples on average, where $H_{\Delta,p^\star} \approx 1/(p^\star\Delta^2)$ is the problem dependent constant. The fixed confidence result of [4] translates, in the fixed budget setting, into an upper bound on the probability of error $\mathrm{e}(T)$ of order $\exp(-c\sqrt{Tp^\star\Delta^2})$ where $c > 0$ is some universal constant – which is much larger than our bound for large T. Similarly, [16] consider the regret with respect to a fixed quantile of order $p^\star$ of the distribution of the means in the reservoir which is again quite related to the regret in our setting. They obtain an algorithm with a bound on cumulative regret of order $R(T) \leq O\big(1/p^\star + \sqrt{(T/p^\star)\log(p^\star T)}\big)$, for any $\Delta > 0$ – in this sense, this analysis is problem independent.

Also closely related is the paper [32] which deals with identifying an ε good arm – in the case where there are many such ε good arms, with high probability. Again this can be seen as a fixed confidence version of our setting, with the proportion of ε good arms being equivalent to our $p^\star$. However, the focus of their results differs considerably to our own. Specifically, in our setting, Theorem 2 of [32] provides an upper bound on the expectation of a stopping time for epsilon good arm identification, of the order $\bar{\mathcal{H}}\log(\bar{\mathcal{H}})$ where $\bar{\mathcal{H}} \approx 1/(p^\star\Delta^2)\log(1/\delta)$ but this bound does not hold in high probability, which would be necessary if one wished to directly compare their results to ours. Indeed for the stopping time of their algorithm to be bounded in high probability one would need to pay a $\log(1/\delta)^2$ term, corresponding to $\exp(-\sqrt{\Delta^2 p^* T})$ in our setting, see Remark 4 in [32] and page 15 in the appendix of the full version [31]. The focus of [32] is instead to get more complete gap dependent bounds, considering also the gaps within the epsilon good arms but as mentioned their results cannot be applied directly to our setting and, as they point out, extending their approach to include high probability guarantees would be strictly sub optimal compared to our results.

We can also view the *most-biased coin problem* studied by [14] and [27] as a particular instance of our setting where all optimal arms are distributed according to a Bernoulli distribution $\mathcal{B}\mathrm{er}(\mu^\star)$ and any sub-optimal arm is distributed according to the *same* Bernoulli distribution $\mathcal{B}\mathrm{er}(\mu^-)$. The goal is then to identify an optimal arm with high probability with as few samples as possible. Precisely, [27] prove that they can find an optimal arm with probability at least $1-\delta$ with $\log\big(1/(p^\star\Delta^2)\big)\frac{\log(1/\delta)}{p^\star\Delta^2}$ samples in expectation when $\mu^\star, \mu^-$ and $p^\star$ are unknown to the agent and with $\frac{\log(1/\delta)}{p^\star\Delta^2}$ samples if $p^\star$ is known. It is also worth mentioning the problem of $p^\star$ estimation for the biased coin problem.

For unknown $p^\star$ and Δ, [39] describe, in the fixed confidence setting, the optimal learning rate for estimating $p^\star$, up to an additive error ε, of the order $\frac{p^\star}{\varepsilon^2\Delta^2}\log(1/\delta)$.

The translation of the result from [27] to the fixed budget setting is much closer to our result, as it would provide a bound of order $\exp\big(-cTp^\star\Delta^2/\log(1/(p^\star\Delta^2))\big)$ where $c > 0$ is some universal constant. This is very similar to our bound, but there is a main difference: we do not assume that there are just two possible distribution for the arms as [27] – the set $\mathcal{A}_{sub}$ of sub-optimal arms might contain arms of diverse means, all being at a gap more than Δ from μ^*. This makes the problem *significantly more difficult* – in particular regarding the adaptation to $p^\star$ – since in our setting, it is impossible to estimate the minimal gap Δ, see Section 5. In fact, extending to a more general reservoir is an open question of interest left at the end of the above paper.

Otherwise, there are some other formulations of the infinitely-many armed bandit problem that are quite popular, but very different from our setting, and that we mention here for completeness. Many works are devoted to the setting where there is some topological relation between the index of the arms, and the mean of the arms [35, 9, 23]. This setting is often referred to as the $\mathcal{X}-$armed bandit setting, and not related to our work as we do not make such topological assumptions. Finally, a paper in which the setting is close to ours, but where the goal is very different, is the one by [28]. The authors consider a partition of the (infinite) space Ω of K-armed bandit models $\nu = (\nu_1, \dots, \nu_K)$, and want to identify for a given bandit model $\mu \in \Omega$ the correct partition component it belongs to.

Fixed confidence to fixed budget setting In the fixed confidence setting for best-arm identification, given some $\delta > 0$, one aims to bound the expected number of samples one needs to correctly identify an optimal arm with probability greater than $1 - \delta$. With our best-arm identification upper bound (Theorem 4) in mind, we can essentially translate our result to the fixed confidence setting by considering $\delta = \exp\left(-\frac{Tp^\star\Delta^2}{\log(1/\Delta)}\right)$, and solving for T. This leads to a upper bound on the number of samples `Elimination` needs to be δ-approximately correct of: $\frac{\log\left(\frac{1}{\delta}\right)\log\left(\frac{1}{\Delta}\right)}{p^\star\Delta^2}$. The papers [27] and [4] both deal with settings very related to our own but from the fixed confidence perspective. [4] deals with quantile estimation and as highlighted above their results can be applied to our setting but with a significantly worse bound on probability of error of order $\exp(\sqrt{Tp^\star\Delta})$. In [27] the problem of best-arm identification is tackled directly but with strong restriction on the reservoir distribution, they consider the case were all sub optimal arms are identically distributed.

Pair matching An additional setting that can be seen in the context of our problem is that of pair matching. Here the learner is presented with a finite graph of nodes, N. The set of nodes N is partitioned into 2 or more communities. The general idea is that nodes in the same community are more likely to be connected by an edge than those in separate communities. A simple and well studied situation is where the graph is generated according to a stochastic block model (SBM), see [25]. In this setting the probability of an edge forming between two nodes of the same community is p and the probability of an edge forming between two nodes of differing communities is q, with $p > q$. Much of the literature is then concerned with identifying communities given complete access to the graph, see [38] and references therein. Of more relation to our specific setting is the paper [22]. Here the learner does not immediately observe the complete graph but is instead able to sequentially query whether two nodes are connected up to a budget T. Their objective is then to minimise their sampling regret, the number of times they query and edge between 2 nodes of differing communities. The problem can be viewed as a bandit problem where each pair of vertices represents an arm following a bernoulli distribution of mean p or q. In our setting the minimal proportion p^* would then be the proportion of pairs which belong to the same community and the gap as $\Delta = p - q$. The fundamental difference is that each arm can only be pulled once, making the problem significantly harder, however, the learner can exploit the SBM structure to their advantage. Assuming the case with exactly two equally sized communities with $T \leq |N|^2$, in [22] they show it is possible to attain a sub linear regret of the order $T\Delta \wedge \frac{(p+q)T}{\Delta}$. The significant worsening of their rate, in comparison to our own, is due to the fact one cannot sample an arm more than once, which significantly changes the flavour of their algorithms.

2 Cumulative regret

We first present an algorithm and prove an upper bound on its cumulative regret, and then we present a problem-dependent lower bound that shows we match the regret bound up to poly-log terms in Δ. Lastly, we provide a theorem to the effect that adaptation to the proportion of optimal arms $p^\star$ is not possible in this setting.

2.1 Upper bound

We present Sampling-UCB for cumulative regret minimization. This algorithm is an Upper Confidence Bound (UCB) type algorithm [37]. We first sample a set $\mathcal{L}$ of arms large enough such that with high probability (of order $1 - 1/T$) there is a proportion of order $p^\star$ optimal arms. Then we build an upper confidence bound on the empirical mean of each sampled arm, see (2), where $\widehat{\mu}_a^t$ is the empirical mean of arm a at time t and N_a^t the number of times arm a was pulled until time t. At time t we pull the arm $a \in \mathcal{L}$ with the highest upper confidence bound U_a^t. The complete procedure is detailed in Algorithm 1. Notably, we do not tune the upper confidence bounds such that they are exceeded with probability less than $1/T$, as for finite-armed bandits. In that setting, a common choice is to have bonuses of the form $\widehat{\mu}_a^t + \sqrt{2\log(T)/N_a^t}$, see [37]. Instead we use an exploration function that does not depend on T, such that the upper confidence bounds are exceeded with probability smaller than a fixed constant, see (2). Thus we only pay a constant regret of order $\log(1/\Delta)$ on the set of sampled arms $\mathcal{L}$. This is made possible by leveraging the fact that we know that there is a proportion of order $p^\star$ optimal arms.

```
**Input:** $\gamma \in (0,1)$, $L \geq 1$
**Initialize:** Pick $\mathcal{L}$, with $|\mathcal{L}| = L$, arms from the reservoir $\mathcal{A}$. Sample each arm once.
**for** $t = L+1$ *to* $T$ **do**
    Compute for each arm $a \in \mathcal{L}$ the quantity
        $$U_a^t = \widehat{\mu}_a^t + \sqrt{\frac{\gamma^2(1-\gamma)^{-1}/4 + \log(\pi^2/6) + 2\log(N_a^t)}{2N_a^t}}, \qquad (2)$$
    Play $a_t = \arg\max_{a\in\mathcal{L}} U_a^t$.
**end**
```

Algorithm 1: Sampling UCB

We prove the following regret bound for Sampling-UCB in Appendix A.

Theorem 1. *For $T \geq 2$, $\gamma \in (0,1)$ and $L = \lceil 4\log(T)/(p^\star\gamma^2)\rceil$, the expected cumulative regret of* Sampling-UCB *is upper bounded as follows:*

$$\mathbb{E}R(T) \leq O\left(\frac{\log(T)\log(1/\Delta)}{p^\star\Delta}\right),$$

see the end of the proof for a precise bound, i.e. (3).

Note that this bound matches the lower bound of Theorem 2 of Section 2.2, for T large enough and up to a $\log(1/\Delta)$ multiplicative factor. Also, L can be calibrated with a lower bound on $p^\star$ instead of $p^\star$, but this lower bound will appear in the rate instead of $p^\star$.

Remark 1. Algorithm Sampling-UCB samples L arms uniformly at random from the reservoir. What we mean by this is that each arm is pulled at random from $\mathcal{A}$ *independently from the other pulled arms*. In other words, by doing this, we potentially artificially create several independent copies of the same arm – which might seem counter-intuitive, but is formally not a problem.
What this anyway implies is that the case $|\mathcal{A}| \leq L$ is not a problem – with this idea of independent copies, we can pull more arms from the reservoir than the number $|\mathcal{A}|$ of arms.

Remark 2. Our algorithm is reminiscent of that of [26], which, as our own, uses a UCB which does not depend on the time horizon, but only on the number of times an arm has been pulled. However, they do so for different reasons, namely to adapt to the infinite time horizon of the fixed confidence setting.

2.2 Lower bound

We can prove an equivalent of the [36] lower bound for finite-armed bandits for our setting. The following theorem is proved in Appendix A.

Theorem 2. *Consider $\Delta \in (0, 1/4)$ and $p^\star \in (0, 1/4]$. For any bandit algorithm, there exists a bandit problem in $\mathfrak{B}_{\Delta,p^\star}$ such that*

$$\mathbb{E}R(T) \geq \min\left(\frac{1}{60}\frac{\max\{\log(\Delta^2 T/16), 0\}}{p^\star \Delta}, \sqrt{T}\right)$$

Note that if we consider the gap Δ and the proportion of optimal arms $p^\star$ as fixed and T large in comparison, i.e. $\Delta \gg \sqrt{1/T}$, then our lower bound is of order $\log(T)/(p^\star\Delta)$. This is the problem-dependent regime that we consider in this paper. On the contrary, if $\Delta \approx \sqrt{1/T}$ then our lower bound is of order $\sqrt{T}$. This is rather the problem-independent regime studied by [16]. We can make a parallel between the lower bound in our setting and the one for finite-armed bandits. Indeed, if we consider that the proxy for the number of arms is $|\mathcal{A}| \sim 1/p^\star$ which implies that there is $p^\star|\mathcal{A}| \sim 1$ optimal arm, then we recover the problem-dependent lower bound of order $|\mathcal{A}|\log(T)/\Delta$, if there are $|\mathcal{A}| - 1$ sub-optimal arms with gap Δ.

2.3 Impossibility of adapting to $p^\star$

The following theorem shows that in the setting of minimizing the cumulative regret, it is impossible to adapt to the proportion of optimal arms $p^\star$. The theorem is proved in Appendix A.

Theorem 3. *Let $p^\star \leq \frac{1}{4}$ and $c > 0$ such that $T \geq 4\left(\frac{c\log(T)}{p^\star\Delta^2}\right)^2$. For any bandit algorithm $\mathfrak{A}$ such that for all bandit problems in $\mathfrak{B}_{\Delta,p^\star}$, we have,*

$$\mathbb{E}R(T) \leq \frac{c\log(T)}{p^\star\Delta}$$

one has that $\forall q^\star \leq \frac{4p^\star}{c}$ there exists a problem in $\mathfrak{B}_{\Delta,q^\star}$ such that

$$\mathbb{E}R(T) \geq \frac{\sqrt{T}\Delta}{4}\ .$$

Remark 3. The `Sampling-UCB` algorithm takes a user defined parameter γ (which can be taken as a universal constant) and L, which should be calibrated depending on (a lower bound on) $p^\star$. While this is necessary, it is important to not that none of the parameters requires knowledge of Δ.

3 Best-arm identification

We present our `Elimination` algorithm for best-arm identification, together with an upper bound on the probability of outputting a sub-optimal arm; next we prove a lower bound, which is matched by our upper bound up to a $1/\log(T)$ factor in the exponential.

3.1 Upper bound

As its name suggests, the `Elimination` algorithm (summarized in Algorithm 2) works by successive elimination of arms – through the update at round i of a set $\mathcal{A}_i$ – although with a twist. We begin by sampling approximately T arms at the first round. Namely, we first select a set $\mathcal{A}_1$ of $|\mathcal{A}_1| = \lfloor \bar{c}T/\log T\rfloor$ arms taken at random from the reservoir, for some constant $\bar{c} > 0$. Then at each round we use a $T/\log T$ fraction of our budget to sample the arms in our set. And so at round i we sample each arm in the set $\mathcal{A}_i$ a number of $t_i = \lfloor \bar{c}T/(|\mathcal{A}_i|\log T)\rfloor$. We then eliminate half of the arms based on the arms' empirical means – namely, we just keep the $\lfloor |\mathcal{A}_i|/2\rfloor \vee 1$ arms in $\mathcal{A}_i$ that have highest empirical means – and introduce an additional number of arms sampled from the reservoir distribution – namely $\lfloor |\mathcal{A}_i|/4\rfloor$ – such that the final size of our arm set is reduced by $\frac{3}{4}$. At the end of the budget, we have one arm remaining – due to the choices of $\bar{c}$ – which is the arm that we return. Note that Remark 1 applies here too so that it is not a problem if $|\mathcal{A}|$ is smaller than the number of arms required by the algorithm. Theorem 4 is proved in Appendix B.

```
Input: $\bar{c}$
set $i \leftarrow 1$
while $i < \log T/\bar{c}$ do
    Sample each arm in $\mathcal{A}_i$ a number $t_i = \lfloor \bar{c}T/(|\mathcal{A}_i| \log T) \rfloor$ of times and compute their
      empirical means $(\hat{\mu}_i(a))_{a \in \mathcal{A}_i}$
    Put in $\mathcal{A}_{i+1}$ the $1 \vee \lfloor |\mathcal{A}_i|/2 \rfloor$ arms that have highest empirical means $(\hat{\mu}_i(a))_{a \in \mathcal{A}_i}$, and add
      on top of that $\lfloor |\mathcal{A}_i|/4 \rfloor$ new arms taken at random from the reservoir
    $i \leftarrow i + 1$
end
Return any $\hat{a}_T$ in $\mathcal{A}_i$
```

Algorithm 2: Elimination

Theorem 4. *Set* $\bar{c} = \log(4/3)$ *.* Elimination *satisfies*

$$\mathbb{P}(\hat{a}_T \in \mathcal{A}^\star) \geq 1 - 2\log(T)\exp\left(-c\frac{\Delta^2 p^\star T}{\log T}\right),$$

where $c = \bar{c}/19200$

Remark 4. Elimination works by discarding many sub-optimal arms and few optimal arms in each round, so that at the end, when just one arm remains, it is optimal with high probability. A key element is that Elimination adds *fresh arms* from the reservoir at each round. This is to ensure that our algorithm is adaptive to $p^\star$, Δ, as ensured by Theorem 4. Whenever the arms in $\mathcal{A}_i$ are pulled less than about Δ^{-2} times, there is no guarantee on what happens when half of the arms are eliminated. Therefore, we have to make sure that when the algorithm arrives at a round i such that $t_i \gtrsim \Delta^{-2}$, the proportion of optimal arms is of larger order than $p^\star$ with high enough probability. This is ensured by adding the fresh arms added from the reservoir. Note that for some arm distributions, we do not need to add fresh arms and the algorithm would function also by just halving at each step the number of arms. Indeed, in the case where all arms follow a Bernoulli distribution, in terms of preserving the proportion of optimal arms, one can prove that halving the set of arms according to the empirical means is no worse than random halving of the set. Thus, in this case, with high probability we increase the proportion of optimal arms at each step, without diminishing it. This is however specific to the case of Bernoulli distributions and some other parametric families, and it is an open question whether this would be true in general.

Remark 5. The successive halving strategy our algorithm for best-arm identification is based on was first introduced by [29], however, without the trick of adding fresh arms, as they didn't need to be adaptive to $p^\star$.

3.2 Lower bound

The following Theorem provides a lower bound on the probability of error for best arm identification in our setting. The proof of Theorem 5 can be found in Appendix B.

Theorem 5. *Consider* $\Delta \in (0, 1/4)$ *and* $p^\star \in [0, 1/4]$*. For any bandit algorithm, there exists a bandit problem in* $\mathfrak{B}_{\Delta,p^\star}$ *such that*

$$\mathrm{e}(T) \geq \frac{1}{4}\exp\left(-Tp^\star\frac{\Delta^2}{32}\right).$$

In proving the above theorem we essentially show that an agent cannot accurately distinguish between two cases: $\mu^* = \frac{1}{2}$ and $\mu^* = \frac{1}{2} + \Delta$. That is, we consider two reservoirs $\mathbf{R}_0$ and $\mathbf{R}_1$ where $\mu_0^* = \frac{1}{2}$ and $\mu_1^* = \frac{1}{2} + \Delta$. Using a coupling argument we bound the KL divergence between the distribution of samples collected on $\mathbf{R}_0$ and $\mathbf{R}_1$. The results then follows by application of Bretagnolle-Huber's inequality.

4 Experiments

We conduct a preliminary set of experiments to test the performance of our algorithms. Specifically, for cumulative regret we compare our Sampling-UCB to the QRM1 algorithm by [16] and the

SR algorithm by Zhu and Nowak [44]. For simple regret we compare our `Elimination` to the BUCB algorithm by [32]. In both cases our performance appears comparable to the literature. See Appendix D for details.

5 Conclusion and open questions

Classifying optimal learning rates on the continuous armed bandit problems with a proportion of optimal arms and general reservoir distribution has been a question of interest in the literature for some time, see [27]. Recent papers – [4] and [44], while focused on a slightly different setting, have considerably weaker results when applied to our setting. Therefore, we believe our results mark a significant improvement in the state of the art. An extension of our results would be to remove the $\log(1/\Delta)$ discrepancy between UB and LB for cumulative regret. However, this appears non-trivial and in particular we struggle to see how a UCB based strategy would achieve this tighter bound in the case of the cumulative regret. Another possibility for further work is an expansion of our setting. Consider the arm reservoir $\mathcal{A}$ partitioned into K possible distributions, each with associated probability p_k. Let $k^* = \arg\max_{[K]} \mu_k$ and take gaps $(\Delta_k)_{[K]} = (\mu_{k^*} - \mu_k)_{[K]}$. One could then consider more detailed bounds, dependent on the sequence $((p_k, \Delta_k))_{[K]}$ as opposed to just $p^\star$ and the smallest gap. The main difficulty here would be to deal with the case where some p_k are much smaller than the proportion $p^\star$ corresponding to the optimal arm.

Acknowledgements The work of J. Cheshire is supported by the Deutsche Forschungsgemeinschaft (DFG) GRK 2297 MathCoRe. The work of P. Ménard is supported by the SFI Sachsen-Anhalt for the project RE-BCI. The work of A. Carpentier is partially supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1), by the DFG - 314838170, GRK 2297 MathCoRe, by the FG DFG, by the DFG CRC 1294 'Data Assimilation', Project A03, by the Forschungsgruppe FOR 5381 „Mathematische Statistik im Informationszeitalter – Statistische Effizienz und rechentechnische Durchführbarkeit", Project 02, by the Agence Nationale de la Recherche (ANR) and the DFG on the French-German PRCI ANR ASCAI CA 1488/4-1 "Aktive und Batch-Segmentierung, Clustering und Seriation: Grundlagen der KI" and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18 and by the SFI Sachsen-Anhalt for the project RE-BCI.

References

[1] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

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

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

[4] Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*, pages 3–24, 2018.

[5] Donald A Berry, Robert W Chen, Alan Zame, David C Heath, and Larry A Shepp. Bandit problems with infinitely many arms. *The Annals of Statistics*, pages 2103–2116, 1997.

[6] Thomas Bonald and Alexandre Proutiere. Two-target algorithms for infinite-armed bandits with bernoulli rewards. *Advances in Neural Information Processing Systems*, 26:2184–2192, 2013.

[7] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[8] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[9] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

[10] Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[11] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[12] Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604, 2016.

[13] Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pages 1133–1141, 2015.

[14] Karthekeyan Chandrasekaran and Richard Karp. Finding a most biased coin with fewest flips. In *Conference on Learning Theory*, pages 394–407, 2014.

[15] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Pac identification of a bandit arm relative to a reward quantile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[16] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Quantile-regret minimisation in infinitely many-armed bandits. In *UAI*, 2018.

[17] Yahel David and Nahum Shimkin. Infinitely many-armed bandits with unknown value distribution. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 307–322. Springer, 2014.

[18] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

[19] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

[20] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29:784–792, 2016.

[21] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

[22] Christophe Giraud, Yann Issartel, Luc Lehéricy, and Matthieu Lerasle. Pair matching: When bandits meet stochastic block model. *stat*, 1050:17, 2019.

[23] Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. *Advances in Neural Information Processing Systems*, 28:667–675, 2015.

[24] Hédi Hadiji. Polynomial cost of adaptation for x-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1029–1038, 2019.

[25] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[26] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[27] Kevin G Jamieson, Daniel Haas, and Benjamin Recht. The power of adaptivity in identifying statistical alternatives. In *Advances in Neural Information Processing Systems*, pages 775–783, 2016.

[28] Sandeep Juneja and Subhashini Krishnasamy. Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory*, pages 1824–1852. PMLR, 2019.

[29] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

[30] Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.

[31] Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. *arXiv preprint arXiv:1906.06594*, 2019.

[32] Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1781–1791. PMLR, 26–28 Aug 2020.

[33] Emilie Kaufmann and Aurélien Garivier. Learning the distribution with largest mean: two bandit frameworks. *ESAIM: Proceedings and surveys*, 60:114–131, 2017.

[34] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[35] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.

[36] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[37] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[38] Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration of random graphs and application to community detection. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 2925–2943. World Scientific, 2018.

[39] Jasper C.H. Lee and Paul Valiant. Uncertainty about uncertainty: Optimal adaptive algorithms for estimating mixtures of unknown coins*. *ACM-SIAM*, 2021.

[40] Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. *31st Annual Conference on Learning Theory*, 75:1–30, 2018.

[41] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

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

[43] Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. *Advances in Neural Information Processing Systems*, 21:1729–1736, 2008.

[44] Yinglun Zhu and Robert Nowak. On regret with multiple best arms. In *Advances in Neural Information Processing Systems*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] In the theorem statements.
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Cumulative regret proofs

A.1 Upper Bound

Proof of Theorem 1. We denote by $\mathcal{L}$ the set of arms sampled from the reservoir such that $|\mathcal{L}| = L$. We also denote by $\mathcal{L}^\star = \{a \in \mathcal{L} : \ a \in \mathcal{A}^\star\}$ the set of optimal arms in $\mathcal{L}$ and by $L^\star = |\mathcal{L}^\star|$ its cardinality. Note that these quantities are all random.

Because of the choice of $L = \lceil 4\log(T)/(p^\star\gamma^2)\rceil$, we know that with high probability there is at least a proportion of $\gamma p^\star$ optimal arms in $\mathcal{L}$. Precisely, if we denote this favorable event by $\mathcal{E} = \{L^\star/L \geq (1-\gamma)p^\star\}$ then by Chernoff's inequality (see Lemma 3), we have

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\big(L^\star/L < (1-\gamma)p^\star\big) \leq e^{-\frac{\gamma^2}{4}Lp^\star} \leq \frac{1}{T}\,.$$

We can decompose the regret given this event and its complement:

$$\begin{aligned}\mathbb{E}[R(T)] &= \mathbb{E}\left[\sum_{a\in\mathcal{L}}(\mu^\star - \mu_a)\mathbb{E}[N_a^T|\mathcal{L}]\mathbb{1}_{\mathcal{E}}\right] + T\mathbb{P}(\mathcal{E}^c)\\ &\leq \mathbb{E}\left[\sum_{a\in\mathcal{L}/\mathcal{L}^\star}\Delta_a\mathbb{E}[N_a^T|\mathcal{L}]\mathbb{1}_{\mathcal{E}}\right] + 1\,.\end{aligned}$$

We now follow the classical proof of UCB-type strategies to upper-bound the number of times a sub-optimal is pulled. From now on, we fix a set of sampled arms $\mathcal{L}$. Fix an $a \in \mathcal{L}\setminus\mathcal{L}^\star$. We have

$$\mathbb{E}[N_a^T|\mathcal{L}] \leq 1 + \sum_{t=L+1}^{T}\mathbb{P}(\forall b\in\mathcal{L}^\star,\, U_{t-1}^b \leq \mu^\star|\mathcal{L}) + \mathbb{P}(a_t = a,\, U_{t-1}^a \geq \mu^\star|\mathcal{L})\,.$$

For the first term in the summation we use the fact that there are many optimal arms. Precisely, using Hoeffding's inequality, we have

$$\begin{aligned}\mathbb{P}(\forall b\in\mathcal{L}^\star,\, U_{t-1}^b \leq \mu^\star|\mathcal{L}) &\leq \mathbb{P}\Bigg(\forall b\in\mathcal{L}^\star, \exists n\in[T] :\ \widehat{\mu}_{b,n}\\ &\qquad\qquad + \sqrt{\frac{\gamma^2(1-\gamma)^{-1}/4 + \log(\pi^2/6) + 2\log(n)}{2n}} \leq \mu^\star\Bigg|\mathcal{L}\Bigg)\\ &\leq \prod_{b\in\mathcal{L}^\star}\left(\sum_{n=1}^{T}\frac{1}{n^2}e^{-\gamma^2(1-\gamma)^{-1}/4-\log(\pi^2/6)}\right)\\ &= e^{-\frac{\gamma^2}{4}(1-\gamma)^{-1}L^\star}\,.\end{aligned}$$

For the second term we proceed as usual. Let

$$n_0 = \inf\left\{n\in\mathbb{N} : \sqrt{\frac{\gamma^2(1-\gamma)^{-1}/4 + \log(\pi^2/6) + 2\log(n)}{2n}} \leq \Delta/2\right\}$$

be such that pulling any arm $a \in \mathcal{A}_{sub}$ more than n_0 times is a small probability event. Note that thanks to Lemma 4

$$n_0 \leq 4\frac{(1-\gamma)^{-1} + \log\big(24(1-\gamma)^{-1}/\Delta^2\big)}{\Delta^2} + 1\,.$$

Then, using again Hoeffding's inequality for an arm $a \in \mathcal{L}\setminus\mathcal{L}^\star$, we obtain

$$\begin{aligned}\sum_{t=L+1}^{T}\mathbb{P}(a_t = a,\, U_{t-1}^a \geq \mu^\star|\mathcal{L}) &\leq \sum_{n=n_a+1}^{T}\mathbb{P}(\widehat{\mu}_{a,n} - \mu \geq \Delta/2) + n_0\\ &\leq \sum_{n\geq 1}e^{-n\Delta^2/2} + n_0 \leq n_0 + \frac{2}{\Delta^2}\,.\end{aligned}$$

Collecting the previous inequalities we can conclude for $T \geq 2$

$$\begin{aligned}
\mathbb{E}[R(T)] &\leq \mathbb{E}\left[\sum_{a \in \mathcal{L}/\mathcal{L}^\star} T e^{-\gamma^2(1-\gamma)^{-1}L^\star/4}\mathbb{1}_{\mathcal{E}} + 1 + \Delta n_0 + \frac{2}{\Delta}\right] + 1 \\
&\leq \mathbb{E}\left[\sum_{a \in \mathcal{L}/\mathcal{L}^\star} T e^{-\gamma^2 L/4}\mathbb{1}_{\mathcal{E}} + 1 + \Delta n_0 + \frac{2}{\Delta}\right] + 1 \\
&\leq L\left(2 + \Delta n_0 + \frac{2}{\Delta}\right) + 1 \\
&\leq \frac{8\log(T)}{p^\star \Delta \gamma^2}\Big(10(1-\gamma)^{-1} + 4\log\left(24(1-\gamma)^{-1}/\Delta^4\right)\Big) + 1\,. \qquad (3)
\end{aligned}$$

□

A.2 Lower Bound

We denote by $\mathcal{B}\mathrm{er}(p)$ the Bernoulli distribution of parameter p. The Kullback-Leibler (KL) divergence between probability distributions P and Q is denoted by $\mathrm{KL}(P, Q)$. In particular, the KL divergence between two Bernoulli distributions $\mathcal{B}\mathrm{er}(p)$ and $\mathcal{B}\mathrm{er}(q)$ is

$$\mathrm{kl}(p,q) = \mathrm{KL}\big(\mathcal{B}\mathrm{er}(p), \mathcal{B}\mathrm{er}(q)\big) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

Proof of Theorem 2. We fix a partition of the reservoir $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3$ and set $p^\star$ the probability to sample an arm in $\mathcal{A}_1$, $\mathcal{A}_2$ and $1 - 2p^\star$ the probability to sample an arm in $\mathcal{A}_3$. We define two bandits problems associated with this reservoir. The bandit problem ν where the arms in $\mathcal{A}_1$ have probability distribution $\mathcal{B}\mathrm{er}(1/2)$, the arm in $\mathcal{A}_2$ and $\mathcal{A}_3$ have probability distribution $\mathcal{B}\mathrm{er}(1/2 - \Delta)$. The second bandit problem ν' is such that the arms in $\mathcal{A}_1$ have probability distribution $\mathcal{B}\mathrm{er}(1/2)$, the arms in $\mathcal{A}_2$ have probability distribution $\mathcal{B}\mathrm{er}(1/2 + \Delta)$ and the arms in $\mathcal{A}_3$ have probability distribution $\mathcal{B}\mathrm{er}(1/2 - \Delta)$. We denote by $\mathbb{E}_\nu$ respectively $\mathbb{E}_{\nu'}$ the expectation under the bandit problem ν respectively ν'.

Let $N_{\mathcal{A}_i}^T = \sum_{t=1}^T \mathbb{1}_{\{a_t \in \mathcal{A}_i\}}$ be the number of times an arm in $\mathcal{A}_i$ is pulled. Note that since the arms in $\mathcal{A}_2$ and $\mathcal{A}_3$ are indistinguishable for the agent in the problem ν, it holds

$$\mathbb{E}_\nu[N_{\mathcal{A}_2}^T] = \frac{p^\star}{1-p^\star}\mathbb{E}_\nu[N_{\mathcal{A}_2}^T + N_{\mathcal{A}_3}^T].$$

Let I^t be the information available by the agent at time t, i.e. the collection of collected rewards and arms pulled. We denote by $\mathbb{P}_\nu^{I^t}$ respectively $\mathbb{P}_{\nu'}^{I^t}$ the distribution of this random variable under the bandit problem ν respectively ν'. Thanks to the chain rule and the above remark we can upper bound the Kullback-Leibler divergence between these two probability distributions

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_\nu^{I^T}, \mathbb{P}_{\nu'}^{I^T})) &= \mathrm{kl}(1/2 - \Delta, 1/2 + \Delta)\mathbb{E}_\nu[N_{\mathcal{A}_2}^T] \\
&= \mathrm{kl}(1/2 - \Delta, 1/2 + \Delta)\frac{p^\star}{1-p^\star}\mathbb{E}_\nu[N_{\mathcal{A}_2}^T + N_{\mathcal{A}_3}^T] \\
&\leq 22p^\star\Delta^2\mathbb{E}_\nu[N_{\mathcal{A}_2}^T + N_{\mathcal{A}_3}^T] = 22p^\star\Delta\mathbb{E}_\nu\big[R(T)\big]\,, \qquad (4)
\end{aligned}$$

where in the last inequality we used that $p^\star \leq 1/4$ and

$$\mathrm{kl}(1/2 - \Delta, 1/2 + \Delta) = 2\Delta\log\left(1 + \frac{2\Delta}{1/2 - \Delta}\right) \leq \frac{4\Delta^2}{1/2 - \Delta} \leq 16\Delta^2.$$

We assume that

$$\mathbb{E}_\nu\big[R(T)\big] = \Delta\big(T - \mathbb{E}_\nu[N_{\mathcal{A}_1}^T]\big) \leq \sqrt{T}, \qquad \mathbb{E}_{\nu'}\big[R(T)\big] = \Delta\mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T] + 2\Delta\mathbb{E}_{\nu'}[N_{\mathcal{A}_3}^T] \leq \sqrt{T},$$

otherwise the result is trivially true. In particular this implies that

$$1 - \sqrt{\frac{1}{\Delta^2 T}} \leq \frac{\mathbb{E}_\nu[N_{\mathcal{A}_1}^T]}{T} \qquad \frac{\mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T]}{T} \leq \sqrt{\frac{1}{\Delta^2 T}}\,. \qquad (5)$$

Using the contraction of the entropy (see Garivier et al. [21]), the inequality $\mathrm{kl}(x,y) \geq x\log(1/y) - \log(2)$ then (5), we obtain

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\nu}^{I^T}, \mathbb{P}_{\nu'}^{I^T}) &\geq \mathrm{kl}\big(\mathbb{E}_{\nu}[N_{\mathcal{A}_1}^T]/T, \mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T]/T\big)\\ &\geq \frac{\mathbb{E}_{\nu}[N_{\mathcal{A}_1}^T]}{T}\log\left(\frac{T}{\mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T]}\right) - \log(2)\\ &\geq \frac{1}{2}\left(1-\sqrt{\frac{1}{\Delta^2 T}}\right)\log(\Delta^2 T) - \log(2)\,.\end{aligned}$$

The previous inequality with the fact that the Kullback-Leibler divergence is positive yields

$$\mathrm{KL}(\mathbb{P}_{\nu}^{I^T}, \mathbb{P}_{\nu'}^{I^T}) \geq \frac{2}{3}\log(\Delta^2 T/16)^+\,. \tag{6}$$

Indeed if $\Delta^2 T/16 \leq 1$ then (6) is trivially true. In the other case we have

$$\begin{aligned}\frac{1}{2}\left(1-\sqrt{\frac{1}{\Delta^2 T}}\right)\log(\Delta^2 T) - \log(2) &\geq \frac{3}{8}\log(\Delta^2 T) - \frac{1}{4}\log(16)\\ &\geq \frac{3}{8}\log(\Delta^2 T/16)\end{aligned}$$

Combining (4) and (6) allows us to conclude

$$\mathbb{E}_{\nu}\big[R(T)\big] \geq \frac{1}{60}\,\frac{\log(\Delta^2 T/16)^+}{p^\star \Delta}.$$

□

A.3 Impossibility of adaptation to $p^\star$

Proof of Theorem 3. Consider $\Delta \in (0, 1/4)$ and the following two definitions of two reservoir distributions:

- The reservoir distribution $\mathbf{R}_0$ characterised by $p_1 = p^\star$ and $p_2 = 1 - p^\star$ and $\nu_1 = \mathcal{B}(1/2)$ and $\nu_2 = \mathcal{B}(1/2 - \Delta)$.
- The reservoir distribution $\mathbf{R}_1$ characterised by $p_1 = q^\star$, $p_2 = p^\star$ and $p_3 = 1 - q^\star - p^\star$ and $\nu_1 = \mathcal{B}(1/2 + \Delta)$ and $\nu_2 = \mathcal{B}(1/2)$ and $\nu_3 = \mathcal{B}(1/2 - \Delta)$.

Note that the Bernoulli distribution is completely characterised by its mean and so we can use the mean to characterise the distribution. Let $\tilde{\mu} = (\tilde{\mu}_j)_{j \leq T}$ be T i.i.d. means corresponding to T i.i.d. distributions sampled according to the reservoir distribution $\mathbf{R}_1$. Note that $\tilde{\mu}_j \in \{1/2 - \Delta, 1/2, 1/2 + \Delta\}$. Write also $\tilde{\mu}' = (\tilde{\mu}'_j)_{j \leq T}$ for the vector of means such that $\tilde{\mu}'_j = \tilde{\mu}_j$ if $\tilde{\mu}'_j \in \{1/2 - \Delta, 1/2\}$, and $\tilde{\mu}'_j = 1/2 - \Delta$ otherwise. Note that then, we have that $(\tilde{\mu}'_j)_{j \leq T}$ are T i.i.d. means corresponding to T i.i.d. distributions sampled according to the reservoir distribution $\mathbf{R}_0$, by definition of $\mathbf{R}_0$. Write $\mathbb{E}_{\mathbf{R}_1}$ for the expectation according to the distribution of $\tilde{\mu}$, i.e. according to $\mathbf{R}_1^{\otimes T}$, and $\mathbb{E}_{\mathbf{R}_0}$ for the expectation according to the distribution of $\tilde{\mu}'$, i.e. according to $\mathbf{R}_0^{\otimes T}$.

Consider an algorithm $\mathfrak{A}$ and a bandit problem involving Bernoulli distributions characterised by a vector of means $m = (m_j)_{j \leq T}$. Write $\mathbb{P}_m^{\mathfrak{A}}$ for the distribution of the samples obtained by the algorithm run on this problem, and $\mathbb{E}_m^{\mathfrak{A}}$ the associated expectation. Consider now another Bernoulli bandit problem characterised by the means $m' = (m'_j)_{j \leq T}$. We have because of the chain rule

$$\mathrm{KL}(\mathbb{P}_{m'}^{\mathfrak{A}}, \mathbb{P}_{m}^{\mathfrak{A}}) = \sum_{j \leq T} \mathbb{E}_{m'}^{\mathfrak{A}}[T_j]\,\mathrm{kl}(m'_j, m_j),$$

where $\mathbb{E}_{m'}^{\mathfrak{A}}$ is the expectation according to problem m' on which algorithm $\mathfrak{A}$ is used, and where T_j is the number of times arm j is sampled at time T.

From our assumption on $\mathfrak{A}$ we have that $\mathbb{E}_{\mathbf{R}_0}[R(T)] \leq \frac{\log(T)}{p^\star \Delta}$. Now, we can obtain

$$\begin{aligned}
\mathrm{KL}(\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathcal{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathcal{A}}_{\tilde{\mu}}) &= \mathrm{KL}(\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) \\
&\leq \mathbb{E}_{\mathbf{R}_1}\left[\mathrm{KL}(\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}})\right] = \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j \leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j]\,\mathrm{kl}(\tilde{\mu}'_j, \tilde{\mu}_j)\right] \\
&\leq \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j \leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j]\frac{\Delta^2}{16}\mathbf{1}\{\tilde{\mu}_j = 1/2 + \Delta\}\right] \\
&= \mathbb{E}_{\mathbf{R}_0}\left[\sum_{j \leq T} \mathbb{E}_{\tilde{\mu}',\mathfrak{A}}[T_j]\frac{\Delta^2}{16}\mathbf{1}\{\tilde{\mu}'_j = 1/2 - \Delta\}\frac{q^\star}{1 - p^\star}\right] \\
&= \frac{q^\star \Delta}{8}\mathbb{E}_{\mathbf{R}_0}[R(T)] \leq \frac{cq^\star}{8p^\star}\log(T) \leq \frac{1}{2}\log(T),
\end{aligned} \tag{7}$$

where the last equality follows since by definition of $\mathbf{R}_0, \mathbf{R}_1$, conditionally on $\tilde{\mu}'_j = 1/2 - \Delta$, the probability that $\tilde{\mu}_j = 1/2 + \Delta$ is $\frac{q^\star}{1-p^\star} \leq 2q^\star$, and otherwise it is 0. And where the final inequality comes from our assumption $p^\star > \frac{cq^\star}{4}$.

Consider the event,

$$E := \left\{\sum_{j \leq T} T_j \mathbf{1}\{\tilde{\mu}'_j = 1/2\} > T/2\right\}.$$

Note that on $\mathbf{R}_0$, we have $\mu^* = \frac{1}{2}$. Thus, on $\mathbf{R}_0$ the event E^C will signify a regret greater than $\frac{T\Delta}{2}$, similarly on $\mathbf{R}_1$ the event E signifies a regret greater than $\frac{T\Delta}{2}$. Thus,

$$E^C \subset \left\{R_{\mathbf{R}_0}(T) \geq \frac{T\Delta}{2}\right\}, \qquad E \subset \left\{R_{\mathbf{R}_1}(T) \geq \frac{T\Delta}{2}\right\}. \tag{8}$$

Where $R_{\mathbf{R}_0}(T)$ and $R_{\mathbf{R}_1}(T)$ denote the regret of the algorithm on $\mathbf{R}_0$ and $\mathbf{R}_1$ respectively. Now from our assumption upon $\mathfrak{A}$ we have that $\mathbb{E}_{\mathbf{R}_0} R(T) \leq \frac{c\log(T)}{p^\star \Delta}$, therefore Equation (8) leads to,

$$\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}\left(E^C\right) \leq \frac{c\log(T)}{p^\star\Delta} \times \frac{2}{T\Delta}. \tag{9}$$

and in addition we also have,

$$\mathbb{E}_{\mathbf{R}_1} R(T) \geq \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) \times \frac{T\Delta}{2}. \tag{10}$$

Now, using the Bretagnolle-Huber's inequality (see Theorem 14.2 by Lattimore and Szepesvári [37]) in combination with (7) we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}(E^C) + \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) &\geq \frac{1}{2}\exp\bigg(-\mathrm{KL}(\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}})\bigg) \\
&\geq \frac{1}{2\sqrt{T}}.
\end{aligned}$$

This result in combination with Equation (9) gives the following,

$$\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) \geq \frac{1}{2\sqrt{T}} - \frac{2c\log(T)}{p^\star T\Delta^2} \geq \frac{1}{4\sqrt{T}} \tag{11}$$

where the final inequality comes from our assumption $T \geq 4\left(\frac{c\log(T)}{p^\star\Delta^2}\right)^2$. Finally our result follows from combination of Equation (9) and Equation (11).

□

B Best-arm identification proofs

B.1 Upper Bound

Proof of Theorem 4. **Proof-specific notations and preliminary considerations.** At round i, write $K_i = |\mathcal{A}_i|$ and write p_i for the proportion of optimal arms in $\mathcal{A}_i$, namely

$$p_i = |\mathcal{A}_i \cap \mathcal{A}^*|/|\mathcal{A}_i|.$$

We also write M_i for the number of optimal arms in $\mathcal{A}_i$ such that $\hat{\mu}_i(a) \geq \mu^* - \Delta/2$, namely

$$M_i = \left|\{a \in \mathcal{A}_i \cap \mathcal{A}^* : \hat{\mu}_i(a) \geq \mu^* - \Delta/2\}\right|,$$

and N_i for the number of sub-optimal arms in $\mathcal{A}_i$ such that $\hat{\mu}_i(a) \geq \mu^* - \Delta/2$, namely

$$N_i = \left|\{a \in \mathcal{A}_i \cap \mathcal{A}_{sub} : \hat{\mu}_i(a) \geq \mu^* - \Delta/2\}\right|.$$

Note that by definition

$$K_{i+1} = \left(1 \vee \left\lfloor \frac{K_i}{2} \right\rfloor\right) + \left\lfloor \frac{K_i}{4} \right\rfloor.$$

Therefore the following bounds holds

$$\left(\left(\frac{3}{4}\right)^i K_1\right) \vee 1 \geq K_i \geq \left(\frac{1}{2}\right)^i K_1 - 4. \tag{12}$$

We write I for the smallest index i such that $K_i = 1$ and will not investigate what happens at rounds $i > I$. By the upper bound (12) on K_i it holds $I \leq \log_{4/3}(K_1) \leq \log_{4/3}(T)$. Note that since $\log_{4/3}(T) = \bar{c}\log T$, the algorithm terminates with a set containing just one arm.

Step 1: Introduction of high-probability events of interest. We define the constant

$$c = \frac{\bar{c}}{10}.$$

We define j^* as the largest j smaller than or equal to I such that

$$K_j \geq cT\Delta^2/(2\log T).$$

Note that such j^* exists since $K_1 \geq \bar{c}T/(2\log T)$, and since $K_I = 1$. We prove below the following upper bound on j^*. Take any round i. Note that for any k, conditionally on $\mathcal{A}_i$, by Hoeffding's inequality, for any $a \in \mathcal{A}_i$

$$\mathbb{P}\left(\left|\hat{\mu}_i(a) - \mu_i(a)\right| \geq \Delta/2 \middle| \mathcal{A}_i\right) \leq 2\exp(-\Delta^2 t_i/2) = q_i, \tag{13}$$

where $\mu_i(a)$ is the true mean associated with arm a. We now state the following technical lemma proved below.

Lemma 1. *Assume that $p^\star \leq 1/2$, and consider $I \geq i \geq j^*$. Under the assumptions of the theorem, we have*

$$q_i^{-1/2} \geq 200 \geq e^2 - 1, \tag{14}$$
$$\Delta^2 t_i/4 \geq \log 2. \tag{15}$$

We define for $i \geq j^*$ and $\bar{p}_i := \left(\frac{p^\star}{6}(5/4)^{i-j^*} \wedge (1/2)\right)$, the event

$$\xi_i = \{p_i > \bar{p}_i\}.$$

Consider from now on $i \geq j^*$.

Step 2: Lower bound on M_i conditional to ξ_i. We have by definition of M_i:

$$M_i = \sum_{a \in \mathcal{A}_i \cap \mathcal{A}^*} \mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \Delta/2\},$$

where by Equation (13), and conditionally on $\mathcal{A}_i$, the $\mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \Delta/2\}$ are independent and dominate stochastically $\mathcal{B}(1-q_i)$, for any $a \in \mathcal{A}_i \cap \mathcal{A}^*$. And so conditionally on $\mathcal{A}_i$, we have that M_i stochastically dominates $\mathcal{B}(K_i p_i, 1-q_i)$. And so by Chernoff's inequality, for any $x \geq \sqrt{q_i}$:

$$\begin{aligned}\mathbb{P}(M_i - p_i K_i(1-q_i) \leq -x p_i K_i | \mathcal{A}_i) &\leq \left[\frac{e^{x/q_i}}{(1+x/q_i)^{1+x/q_i}}\right]^{K_i p_i q_i} \\ &\leq \exp\Big[x K_i p_i - \log(1+x/q_i)(K_i p_i q_i + x K_i p_i)\Big] \\ &\leq (1+x/q_i)^{-x K_i p_i/2}.\end{aligned}$$

as for $i > j^*$ we have $\log(1+x/q_i) > 2$, see Lemma 1.

So that for $x \geq \sqrt{q_i}$

$$\mathbb{P}(M_i \leq K_i p_i(1-2x)|\mathcal{A}_i) \leq \exp\Big(-x\Delta^2 t_i K_i p_i/16\Big),$$

since $\log(q_i^{-1}) = \Delta^2 t_i/2 - \log 2 \geq \Delta^2 t_i/4$ for $I \geq i \geq j^*$ - see Lemma 1.

And so since $p_i \geq \frac{p^\star}{6}$ on ξ_i

$$\mathbb{P}(M_i \leq p_i K_i(1-2x)|\xi_i) \leq \exp\big(-\bar{c}' x p^\star \Delta^2 T/\log T\big) := u. \tag{16}$$

where $\bar{c}' = \bar{c}/96$ and recalling $t_i = \lfloor \bar{c}T/\big(K_i \log(T)\big)\rfloor$.

Step 3: Upper bound on N_i conditional to ξ_i. We have by definition of N_i:

$$N_i = \sum_{a \in \mathcal{A}_i \cap \mathcal{A}_{sub}} \mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \Delta/2\},$$

where by Equation (13), and conditionally on $\mathcal{A}_i$, the $\mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \bar{\Delta}/2\}$ are independent and are stochastically dominated by $\mathcal{B}(q_i)$, for any $a \in \mathcal{A}_i \cap \mathcal{A}_{sub}$. And so conditionally on $\mathcal{A}_i$, we have that N_i is stochastically dominated by $\mathcal{B}(K_i, q_i)$. And so by Chernoff's inequality for any $x \geq 2$:

$$\mathbb{P}(N_i - K_i q_i \geq x K_i | \xi_i) \leq \left[\frac{e^{x/q_i}}{(1+x/q_i)^{1+x/q_i}}\right]^{K_i q_i} \leq (1+x/q_i)^{-x K_i/2},$$

similar to Step 2.

So that for $x \geq \sqrt{q_i}$

$$\mathbb{P}(N_i \geq 2K_i x|\mathcal{A}_i) \leq \exp\Big(-x\Delta^2 t_i K_i/16\Big),$$

as in Step 2.

And so similar to in Step 2:

$$\mathbb{P}(N_i \geq 2x K_i|\xi_i) \leq \exp\big(-\bar{c}' x \Delta^2 T/\log T\big) \leq u. \tag{17}$$

Step 4: Bound on the probability of ξ_i and conclusion. First we have – since we add $K_{j^*-1}/4 = K_{j^*}/3$ fresh arms to the set $\mathcal{A}_{j^*}$ - that

$$\left\{\left|\sum_{a \in \mathcal{A}_{j^*}} \mathbf{1}\{a \in \mathcal{A}^*\} - \frac{1}{3}p^\star K_{j^*}\right| \leq \frac{1}{6}p^\star K_{j^*}\right\} \subset \xi_{j^*},$$

where it holds that $\mathbf{1}\{a \in \mathcal{A}^*\} \sim \mathcal{B}(p^*)$ for the fresh arms and $|\mathcal{A}_{j^*}| = K_{j^*}$. And so by Chernoff's inequality:

$$\mathbb{P}(\xi_{j^*}) \geq 1 - 2\exp(-p^\star K_{j^*}/10) \geq 1 - 2\exp\left(-c\frac{p^\star T\Delta^2}{20\log T}\right) =: 1 - v, \tag{18}$$

by definition of j^*.

Now consider $i > j^*$, let,

$$\xi_i^{'} = \left\{p_{i+1} \geq \frac{5}{4}p_i \wedge \frac{1}{2}\right\}.$$

Lemma 2. *Assume that* $2x \leq 1/100$. *We have for* $I \geq i > j^*$*:*

$$\xi_i'' := \{M_i > p_i K_i(1-2x)\} \cap \{N_i < 2xK_i\} \subset \xi_i'.$$

Note also that

$$\mathbb{P}(\xi_i''|\xi_i) \geq 1 - 2u,$$

by Equations (16) and (17), so that by Lemma 2

$$\mathbb{P}(\xi_i'|\xi_i) \geq 1 - 2u. \tag{19}$$

By induction it holds that for any $1 \leq m \leq I - j^*$

$$\xi_{j^*} \cap \bigcap_{j^* < i \leq j^*+m} \xi_i' \subset \bigcap_{j^* \leq i \leq j^*+m} \xi_i,$$

so that by Equations (18) and (19)

$$\mathbb{P}\left(\bigcap_{j^* \leq i \leq j^*+m} \xi_i\right) \geq (1-v)(1-2u)^m \geq 1 - v - 2um.$$

In particular using the previous inequality for $m = I - j^*$ and since $I \leq \log T$ it holds

$$\mathbb{P}\left(\bigcap_{j^* \leq i \leq I} \xi_i\right) \geq 1 - v - 2u\log T.$$

Since $K_I = 1$, and since by definition of the ξ_i we know that on ξ_I we have that the only arm in $\mathcal{A}_I$ is optimal, this concludes the proof - taking $x = 1/200$, which is compatible with $x \geq \sqrt{q_i}$ as $q_i \leq 1/200^2$ by Lemma 1.

□

We prove now successively, Lemma 1, Lemma 2 used in the proof of Theorem 4.

Proof of Lemma 1. Note first that for $I \geq i \geq j^*$ we have

$$K_{i+1} = \lfloor K_i/2 \rfloor \vee 1 + \lfloor K_i/4 \rfloor \leq \frac{3K_i}{4} \vee 1.$$

So that for any $0 \leq m < I - j^*$ we have by definition of I as the first index such that $K_I = 1$

$$K_i \leq K_{j^*}(3/4)^m. \tag{20}$$

Also for any i such that $K_i \geq 4$

$$K_{i+1} \geq K_i/2,$$

and for any i such that $K_i < 4$, we have

$$K_{i+1} = 1,$$

so that for any $0 \leq m < I - j^*$ we have

$$K_i \geq K_{j^*}(i/2)^m.$$

Inequality (14): We therefore have for $I > i \geq j^*$ and by Equation (20)

$$\begin{aligned} q_i^{-1/2} = 2^{-1/2}\exp(\Delta^2 t_i/4) &\geq 2^{-1/2}\exp\left(\bar{c}\frac{\Delta^2 T}{2K_{j^*}\log(T)}\right), \\ &\geq 2^{-1/2}\exp(10) \geq 200 \geq e^2 - 1 \end{aligned}$$

Inequality (15): We have,

$$q_i = \exp(-\Delta^2 t_i/2) \ ,$$

thus by inequality (14) we have

$$\exp(\Delta^2 t_i/4) \geq \sqrt{2}(e^2 - 1),$$

so that

$$\Delta^2 t_i/4 \geq \log 2.$$

□

Proof of Lemma 2. Let i such that $I \geq i > j^*$. Note that on $\xi_i{}''$, we have $M_i > 0$ so that $p_i > 0$.

First case: $0 < p_i \leq 2/5$**.** Assume first that $p_i \leq 2/5$. On ξ_i'' we have that

$$M_i > p_i K_i(1 - 2x),$$

and

$$N_i < 2K_i x,$$

so that

$$M_i + N_i < p_i K_i + 2K_i x \leq (2/5)K_i + K_i/100 \leq K_i/2.$$

since $2x \leq 1/100$ for $i \geq j^*$ - see Lemma 1. And so all M_i arms of $\{a \in \mathcal{A}_i \cap \mathcal{A}^* : \hat{\mu}_i(a) \geq \mu^* - \bar{\Delta}/2\}$ are going to be in $\mathcal{A}_{i+1}$. This implies – as in this case $K_i \geq 2$ otherwise we cannot have $0 < p_i \leq 2/5$ – that

$$p_{i+1} \geq \frac{M_i}{K_{i+1}} = \frac{M_i}{1 \vee \lfloor K_i/2 \rfloor + \lfloor K_i/4 \rfloor} \geq \frac{4}{3}(1 - 2x)p_i > \frac{5}{4}p_i,$$

as $2x \leq 1/100$.

Second case: $p_i > 2/5$**.** Assume now that $p_i > 2/5$. On ξ_i'' we have that

$$M_i > p_i K_i(1 - 2x) \geq \frac{198}{500}K_i,$$

and

$$N_i < 2K_i x \leq K_i/100,$$

since $2x \leq 1/100$ for $I \geq i > j^*$ – see Lemma 1. Since $198/500 + 1/100 = 203/500 < 1/2$ this implies that at least $\frac{199}{500}K_i$ from the arms in $\{a \in \mathcal{A}_i \cap \mathcal{A}^* : \hat{\mu}_i(a) \geq \mu^* - \bar{\Delta}/2\}$ are going to be in $\mathcal{A}_{i+1}$. So that

$$p_{i+1} \geq \frac{M_i}{K_{i+1}} = \frac{M_i}{1 \vee \lfloor K_i/2 \rfloor + \lfloor K_i/4 \rfloor} \geq \frac{4}{3} \times \frac{198}{500} = \frac{66}{125} > 1/2.$$

This concludes the proof. □

B.2 Lower Bound

Proof of Theorem 5. We consider a similar setting to that in the proof of Theorem 3 although with a slightly different construction of $\mathbf{R}_0, \mathbf{R}_1$.

Consider the following two reservoir distributions:

- The reservoir distribution $\mathbf{R}_0$ characterised by $p_1 = p^\star$ and $p_2 = 1 - p^\star$ and $\nu_1 = \mathcal{B}(1/2)$ and $\nu_2 = \mathcal{B}(1/2 - \Delta)$.
- The reservoir distribution $\mathbf{R}_1$ characterised by $p_1 = p^\star$ and $p_2 = p^\star$ and $p_3 = 1 - 2p^\star$ and $\nu_1 = \mathcal{B}(1/2 + \Delta)$ and $\nu_2 = \mathcal{B}(1/2)$ and $\nu_3 = \mathcal{B}(1/2 - \Delta)$.

We define $\tilde{\mu}, \tilde{\mu}'$, and associated expectations and probabilities as in the proof of Theorem 3. Consider also any algorithm $\mathfrak{A}$. We have by similar calculations as Equation (7) the following upper bound on the KL divergence

$$\begin{aligned}
\mathrm{KL}(\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) &= \mathrm{KL}(\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) \\
&\leq \mathbb{E}_{\mathbf{R}_1}\left[\mathrm{KL}(\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}})\right] = \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j\leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j]\,\mathrm{kl}(\tilde{\mu}'_j, \tilde{\mu}_j)\right] \\
&\leq \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j\leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j]\frac{\Delta^2}{16}\mathbf{1}\{\tilde{\mu}_j = 1/2+\Delta\}\right] \\
&= \mathbb{E}_{\mathbf{R}_0}\left[\sum_{j\leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j]\frac{\Delta^2}{16}\mathbf{1}\{\tilde{\mu}'_j = 1/2-\Delta\}\frac{p^\star}{1-p^\star}\right], \qquad (21)
\end{aligned}$$

since by definition of $\mathbf{R}_0, \mathbf{R}_1$, conditionally on $\tilde{\mu}'_j = 1/2 - \Delta$, the probability that $\tilde{\mu}_j = 1/2 + \Delta$ is $\frac{p^\star}{1-p^\star}$, and otherwise it is 0.

By Equation (21) and since $\sum_{j\leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j] = T$, we have

$$\mathrm{KL}(\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) \leq T\frac{\Delta^2}{16}\frac{p^\star}{1-p^\star}.$$

Now by Bretagnolle-Huber's inequality (see Theorem 14.2 by Lattimore and Szepesvári [37]) and for any event E

$$\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) + \mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}(E^C) \geq \frac{1}{2}\exp\Bigg(-\mathrm{KL}(\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}})\Bigg). \qquad (22)$$

Let us write $\hat{a}_T$ for the arm that the algorithm $\mathfrak{A}$ recommends. Set

$$E = \{\tilde{\mu}_{\hat{a}_T} = 1/2\}.$$

Note that on E, we make a mistake in prediction for $\tilde{\mu}$, and that on E^C, we make a mistake in prediction for $\tilde{\mu}'$. We have

$$\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) + \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}(E^C) \geq \frac{1}{2}\exp\left(-T\frac{\Delta^2}{16}\frac{p^\star}{1-p^\star}\right).$$

This concludes the proof by definition of E. □

C Technical lemmas

Lemma 3. *(Chernoff bound) Let $X_1, \dots, \mathcal{X}_n \sim \mathcal{B}\mathrm{er}(p)$ be n samples from a Bernoulli distribution and $S_n = \sum_{k=1}^n X_n$ their sum. Then for all $\gamma \in [0,1]$ it holds*

$$\mathbb{P}\left(\frac{S_n}{n} \leq (1-\gamma)p\right) \leq e^{-\frac{\gamma^2}{4}np}\,,$$

$$\mathbb{P}\left(\frac{S_n}{n} \geq (1+\gamma)p\right) \leq e^{-\frac{\gamma^2}{4}np}\,.$$

Proof. We prove the first inequality; the second one is similar. If $(1-\gamma)p < 0$ or $\gamma = 0$ the inequality is trivially true. Else, because of Chernoff's inequality, we have

$$\mathbb{P}\left(\frac{S_n}{n} \leq (1-\gamma)p\right) \leq e^{-n\,\mathrm{kl}\left((1-\gamma)p,p\right)}\,.$$

It remains to remark to conclude that

$$\mathrm{kl}\left((1-\gamma)p,p\right) \geq \frac{\gamma^2}{2}p\,,$$

where we used the refined Pinsker inequality from Garivier et al. [21], for $0 \leq x < y \leq 1$,

$$\mathrm{kl}(y, x) \geq \frac{1}{2\max_{x\leq q\leq y} q(1-q)}(x-y)^2 \geq \frac{1}{2y}(x-y)^2 \,.$$

For the second inequality we use

$$\mathrm{kl}\big((1+\gamma)p, p\big) \geq \frac{1}{2(1+\gamma)p}\gamma^2 p^2 \geq \frac{\gamma^2}{4} p \,.$$

□

Lemma 4. *Let $A, B, C \geq 0$ be constants such that $A \geq C$, then for $n_0 = \inf\{n \geq 1 : \ A + B\log(n) \leq nC\}$ we have*

$$n \leq \frac{A + B\log\big((2(B^2 + AC)/C^2\big)}{C} + 1 \,.$$

Proof. First let $x_0 \geq 1$ be such that $A + B\log(x_0) = Cx_0$. It exists since $A + B\log(x)/x \to 0$ if $x \to \infty$ and since $A \geq C$. In particular, because of the definition of n_0 we have $x_0 \leq n_0 \leq x_0 + 1$. Then note that $A + B\sqrt{x_0} \leq Cx_0$. Thus $\sqrt{x_0}$ is smaller than the largest roots of the polynomial $Cy^2 - By - A$. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $(a+b)^2 \leq 2(a^2 + b^2)$ we obtain

$$\begin{aligned} x_0 &\leq \left(\frac{B + \sqrt{B^2 + 4AC}}{2C}\right)^2 \\ &\leq 2\frac{B^2 + AC}{C^2} \,. \end{aligned}$$

Inserting the previous inequality in the definition of x_0 and using $n_0 \leq x_0 + 1$ allows us to conclude

$$n_0 \leq \frac{A + B\log\big(2(B^2 + AC)/C^2\big)}{C} + 1 \,.$$

□

D Experiments

In this section we conduct preliminary experiments for the cumulative regret and best-arm identification setting.

Cumulative regret For the cumulative regret we compare `Sampling-UCB` (with $\gamma = 0.5$) with the QRM1 algorithm by [16] and SR algorithm by [44]. We arbitrarily[4] choose the following reservoir: the arms are distributed according to a Bernoulli distribution with possible means $[0.5, \ 0.8]$ sampled with probabilities $[0.8, \ 0.2]$. We remark that the SR algorithm and `Sampling-UCB` are very similar, they both sample approximately $\log(T)/p^\star$ arms and run a regret minimizer algorithm on this set of arms. The only difference is that the SR algorithm relies on the MOSS algorithm. Whereas the QRM1 algorithm proceeds by progressively adding new arms. In particular this algorithm is anytime. In Figure 1 we compare the cumulative regret of the different algorithms for a fixed horizon $T = 20000$. We observe that `Sampling-UCB` behaves similarly to SR and that QRM1 performs slightly worst (maybe because of the adaptation to T). We also check that all algorithms have a regret that is logarithmic with the horizon as expected. To this aim, in Figure 2, we plot the cumulative regret (for the same reservoir) for all horizons $T \in \{100, 200, \ldots, 10000\}$.

Best-arm identification For best arm identification we compare our algorithm with the BUCB algorithm by [32]. In Figure 3 we compare the performance of the algorithms across varying Δ for a fixed $T = 1000$. That is, we consider reservoirs of the form $[0.2, \Delta, 1]$ for $\Delta \in (0.01 \times i)_{i\in[79]}$ with probabilities $[0.29, 0.69, 0.02]$. The BUCB algorithm presents an issue as it is designed for the fixed confidence regime the algorithm takes δ as a parameter. We set δ equal to an arbitrarily low constant. The BUCB algorithm works by opening successively large brackets of arms, however as they do not provide results in high probability, only in expectation, they can draw significantly less arms from the reservoir. The performance of `Elimination` seems favourable compared to BUCB, however, one may be able to improve the performance of BUCB with parameter tuning.

[4] Which is not very important, since we evaluate the algorithms from a problem-dependent point of view

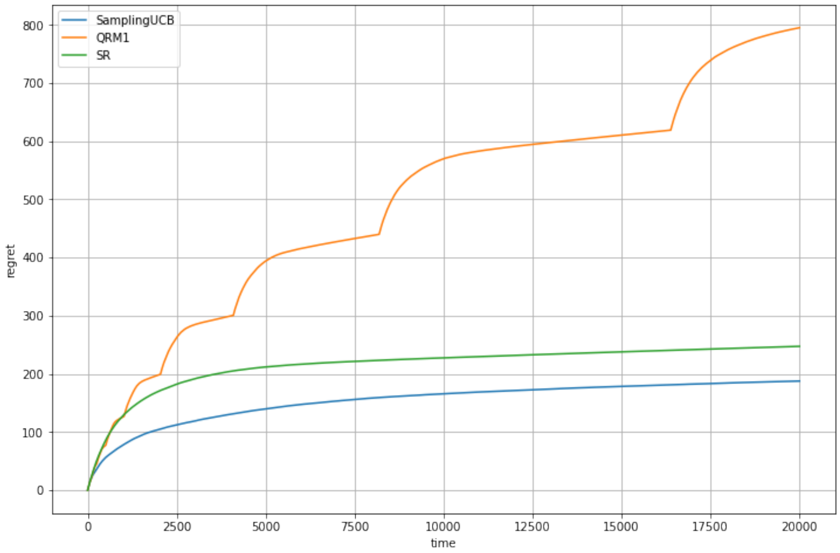

Figure 1: Cumulative regret in function of the time estimated by 100 Monte-Carlo simulations.

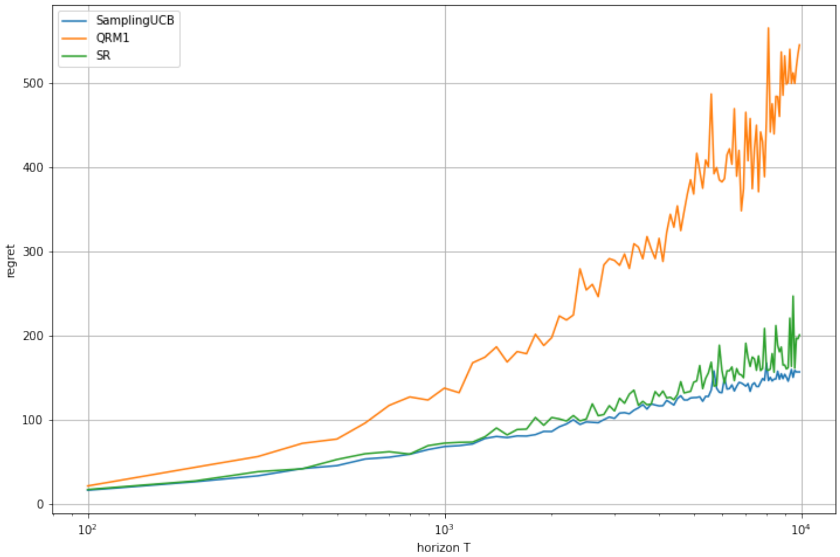

Figure 2: Cumulative regret in function of the horizon $T \in \{100, 200, \dots, 10000\}$ estimated by 100 Monte-Carlo simulations.

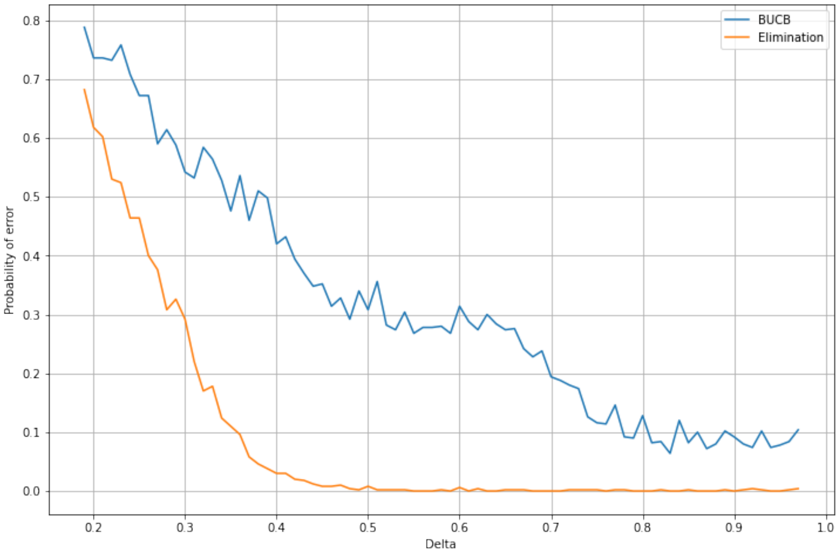

Figure 3: Probability of error for best arm identification across varying Δ using 500 Monte-Carlo simulations.