Proceedings of Machine Learning Research vol 75:1–33, 2018 31st Annual Conference on Learning Theory

Best of both worlds: Stochastic & adversarial best-arm identification

Yasin Abbasi-Yadkori ABBASIYA@ADOBE.COM
Adobe Research, USA

Peter L. Bartlett PETER@BERKELEY.EDU
University of California, Berkeley, USA

Victor Gabillon VICTOR.GABILLON@QUT.EDU.AU
Queensland University of Technology - ACEMS, Australia

Alan Malek AMALEK@MIT.EDU
Massachusetts Institute of Technology, USA

Michal Valko MICHAL.VALKO@INRIA.FR
SequeL team, INRIA Lille - Nord Europe, France

Editors: Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet

Abstract

We study bandit best-arm identification with arbitrary and potentially adversarial rewards. A simple random uniform learner obtains the optimal rate of error in the adversarial scenario. However, this type of strategy is suboptimal when the rewards are sampled stochastically. Therefore, we ask: *Can we design a learner that performs optimally in both the stochastic and adversarial problems while not being aware of the nature of the rewards?* First, we show that designing such a learner is impossible in general. In particular, to be robust to adversarial rewards, we can only guarantee optimal rates of error on a subset of the stochastic problems. We give a lower bound that characterizes the optimal rate in stochastic problems if the strategy is constrained to be robust to adversarial rewards. Finally, we design a simple parameter-free algorithm and show that its probability of error matches (up to log factors) the lower bound in stochastic problems, and it is also robust to adversarial ones.

Keywords: multi-armed bandits, best-arm identification, adversarial and stochastic rewards

1. Introduction

In best-arm identification (Maron and Moore, 1993; Bubeck et al., 2009), the *learner* tries to identify the *arm* (option, decision) with the highest (expected) average *reward* among K given arms. At each round t of the *game* (the interaction of the learner with its environment), each arm k is assigned a reward $\boldsymbol{g}_{k,t}$. On this very same round t, a learner chooses an arm I_t and *only* observes the reward of that arm, $\boldsymbol{g}_{I_t,t}$, while the rest of the vector $\boldsymbol{g}_t$ is hidden from the learner.

Typically, we assume stochastic rewards: Each arm k is associated with a distribution ν_k and, for all k and t, the $\boldsymbol{g}_{k,t}$ is sampled i.i.d. from ν_k. In this paper, we aim for a robust solution for this setting and we allow the possibility that the rewards are non-stochastic. The rewards could have been chosen even by an oblivious *adversary*: The rewards are fixed before the start of the game but they are not necessarily drawn i.i.d. from a distribution. We focus on *fixed-budget*, where a total number of arm pulls n is fixed and the learner wishes to identify, as accurately as possible, the arm

that attains the highest cumulative reward. However, the results extend to a fixed-confidence case as we discuss in Section 6, together with *adaptive* adversaries and other cases.

Given $\boldsymbol{g}$ and round n, we define the cumulative gain of arm k as $G_k = \sum_{t=1}^{n} \boldsymbol{g}_{k,t}$. Same as for adversarial bandits for cumulative-regret (Auer et al., 2002), the best arm in hindsight is defined as $k^\star_{\boldsymbol{g}} = \operatorname{argmax}_{k\in[K]} G_k$.[1] In a similar way, we define the gaps in hindsight with respect to $\boldsymbol{g}$ between two arms k and j as $\Delta^{\boldsymbol{g}}_{k,j} \triangleq \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{g}_{k,t} - \boldsymbol{g}_{j,t})$, giving a good proxy for the difficulty of discriminating between these two arms even in an adversarial environment. We design a learner and show that the probability[2] of error at round n *given* $\boldsymbol{g}$ against any fixed adversarial reward design $\boldsymbol{g}$ is bounded by a measure of complexity depending on the gaps in hindsight with respect to $\boldsymbol{g}$ and n.

Next, we discuss the motivations for studying non-stochastic best-arm identification. First, learning in the presence of adversarial data implies robustness. In a real-world best-arm identification, such as clinical trials or online ad recommendation, the assumption of i.i.d. data may not be valid. For instance, there could a correlation between subsequent pulls of an arm. It is also possible that an adversary is trying to obscure the correct results: For example, the adversary might use a botnet to make the learner sell more ads. As discussed by Bubeck and Cesa-Bianchi (2012, Section 3), a deterministic learner or a learner that eliminates arms with low observed cumulative reward in an early stage of the game could be easily fooled by an adversary feeding it uninformative rewards in each of its deterministic pulls or in early stages of the game. Therefore, an efficient learner needs to employ internal randomization and pull each arm with a positive probability $\boldsymbol{p}_{k,t} \triangleq \mathbb{P}(I_t = k)$.

Best-arm identification with completely adversarial rewards is so difficult that a very conservative approach is already near-optimal. In Section 3 we show that by playing the arms *uniformly at random*, the learner obtains the optimal gap-dependent rates of error against worst-case adversarial sequence. In the stochastic case, however, picking arms uniformly at random is suboptimal. This reveals the **best of both worlds (BOB)** question: *Is there a learner that attains the optimal error rates in both the stochastic and adversarial settings without knowledge of the environment?*

To study the above question, we face a number of **challenges**: How to efficiently mix the needs for randomization and exploration with the urge to pull the most promising arms? Can the learner detect if the rewards are stochastic or adversarial? What is an appropriate estimate of G_k? In the stochastic case, $\widehat{G}_k \triangleq \frac{n\sum_{t=1}^{n}\mathbf{1}\{I_t=k\}\boldsymbol{g}_{k,t}}{\sum_{t'=1}^{n}\mathbf{1}\{I_{t'}=k\}}$ is commonly used, but if not used carefully, it can be easily biased by an adversary. In the adversarial case, $\widetilde{G}_k \triangleq \sum_{t=1}^{n}\frac{\boldsymbol{g}_{k,t}}{\boldsymbol{p}_{k,t}}\mathbf{1}\{I_t = k\}$ is usual but it can have a high variance if $\boldsymbol{p}_{k,t}$ is small (scaling with $\sum_{t=1}^{n}(1/\boldsymbol{p}_{k,t})$). Controlling this potentially large variance, especially when dealing with a stochastic problem, is one of the main challenges that we face. In particular, high variance can happen if a learner explores uniformly for too long.

Our contributions We consider a new formulation of the adversarial best-arm identification. We study whether a BOB result is possible for fixed-budget best-arm identification. We answer the question negatively and show that for a class of stochastic problems, no robust (to adversary) learner can achieve the optimal stochastic error rates. To prove this result, we introduce a new measure of complexity of the stochastic part of the task. This measure of complexity gives the problem-dependent error rates that a robust learner can guarantee in any stochastic problem. We study this new complexity in different stochastic regimes and provide several positive examples where BOB is possible but also several negative ones. We notice that even in stochastic problems where our lower

1. An alternative definition of best-arm identification could be predicting $\operatorname{argmax}_{k\in[K]}\boldsymbol{g}_{k,n+1}$. However, this is impossible in the adversarial case where $\boldsymbol{g}_{k,n+1}$ can be chosen arbitrarily without any dependence on $\boldsymbol{g}_{k,[n]}$.
2. Note that in our setup, given $\boldsymbol{g}$, the randomness comes solely from the potential internal randomization of the learner.

bound seems to indicate that BOB is achievable, the robust uniform learner is clearly suboptimal. Therefore there is a need for a new algorithm to bridge the gap. In Section 5, we design a simple parameter-free learner and show that its error rate matches, up to log factors, the lower bound in the stochastic case as well as those of the uniform strategy in the adversarial case.

Related work The stochastic best-arm identification was introduced in the fixed-budget setting by Bubeck et al. (2009) and Audibert et al. (2010). Refined upper and lower bounds can be found respectively in the works of Karnin et al. (2013) and Carpentier and Locatelli (2016). For cumulative regret, the BOB question was raised by Bubeck and Slivkins (2012). Seldin and Slivkins (2014) gave a practical algorithm for addressing the same problem (see also Seldin and Lugosi, 2017). A lower bound and a refined upper bound for the problem was given by Auer and Chiang (2016).

Best-arm identification has been studied in a different non-stochastic setting by Jamieson and Talwalkar (2016) and Li et al. (2016). At round t for its m-th pull of arm k, their learner observe $\boldsymbol{g}_{k,m}$, whereas our learner observes $\boldsymbol{g}_{k,t}$. Moreover, their work is specifically tailored for online hyperparameter optimization of learning methods, where the value of each hyperparameter is assumed to converge, at some unknown rate, to its true value as it is given more resources (e.g., data or time). Therefore, their objective is to identify the best arm once the convergence has happened while in our setting, we do not assume any convergence and are interested instead in comparing the cumulative rewards G_k of each arm. By assuming convergence of the arms and asymptotically having the learner observe all the rewards, they prove that a state-of-the-art deterministic learner from the vanilla stochastic best-arm identification also solves their setting.

Allesiardo and Féraud (2017) and Allesiardo et al. (2017) analyze a non-stationary stochastic best-arm problem in a fixed-confidence setting but under the assumption that the game can be split into independent sub-games where the identity of the best arm does not change. This assumption precludes many of the hard examples where the adversary tricks the learner in the early stages of the game. The learner can then simply use a randomized version of Hoeffding Races and safely stop pulling arms. Also, while our notion of gaps is defined for round n, they define gaps at any intermediate round of the game but then consider the minimum gaps over time for each arm. This leads to a larger notion of complexity and permits ignoring the variance of the estimates.

Two types of applications In the first type, the learner believes that future rewards will be similar to the ones already observed. Here, from the n observations, the learner could use the estimated best arm in the upcoming rounds. In an expert setting, we might test the quality of experts for a limited time before committing to one. We may also optimize the hyperparameters of algorithms. Contrary to Jamieson and Talwalkar (2016), our objective is to find the arm (hyperparameter) with the highest cumulative value over a test set, rather than the performance that is achieved by the hyperparameter after convergence. In the second type, there are no new upcoming rounds. The identity of the best arm is used to take an action based on the collected data. As an example, consider a law-enforcing agency that collects information periodically from different targets, one target per week, and at the end of the year, it decides which target to investigate thoroughly over its activities in the past year. This problem can be seen as a game between the agency and the malignant targets. Therefore, it would be useful to have algorithms that are robust to worst-case scenarios but still take advantage of easy data in case malignant targets actually do not take precautions to avoid being caught. Finally, in our setting, we are not trying to be robust to corruption of the data, as we want to find the best arm whether it includes corruption or not, unlike Altschuler et al. (2018) who study corrupted bandits.

2. More on the problem formulation in a general adversarial and stochastic settings

Let $[a:b] = \{a, a+1, \dots, b\}$ with $a, b \in \mathbb{N}$, $a \leq b$, and $[a] = [1:a]$. A vector indexed by both round t and a specific element index k is $\boldsymbol{w}_{k,t}$. We detail the general game protocol in Figure 1.

Input: the time horizon n and the number of arms K
For $t = 1, 2, \dots, n$,
- simultaneously, the learner chooses arm $I_t \in [K]$, and
- the adversary picks a gain vector. $\boldsymbol{g}_t \in [0,1]^K$.
- Then, the learner observes $\boldsymbol{g}_{t,I_t}$.

After round n is over, the learner recommends the arm J_n with the objective that $J_n = k^\star_{\boldsymbol{g}}$.

Figure 1: General protocol of the adversarial setting. The adversary is oblivious.

Adversarial case The adversary is denoted as ADV. It is the process that generates $\boldsymbol{g}$. Let (m) denote the index of the m-th best arm in $[K]$ and $G_{(m)}$ its corresponding cumulative gain so that $G_{(1)} > G_{(2)} \geq \ldots \geq G_{(K)}$. Dually to $(\cdot)$, $\langle k \rangle$ denotes the rank of arm k after sorting according to $G_{(\cdot)}$ so that $\langle (k) \rangle = (\langle k \rangle) = k$, $\forall k \in [K]$. Without loss of generality, note that we assumed there exists a unique best arm $k^\star_{\boldsymbol{g}} = (1)$. For each arm $k \in [K]$, we define the gap $\Delta^{\boldsymbol{g}}_k$ as

$$n\Delta^{\boldsymbol{g}}_k \triangleq \begin{cases} G_{(1)} - G_k & \text{if } k \neq k^\star_{\boldsymbol{g}}, \\ G_{(1)} - G_{(2)} & \text{if } k = k^\star_{\boldsymbol{g}}. \end{cases}$$

The gap can also be written as $n\Delta^{\boldsymbol{g}}_k = \left| \max_{i \neq k} G_i - G_k \right|$. $J_n \in [K]$ is the arm returned by the learner at the end of the exploration phase. Given a budget n and a fixed adversarial set of rewards $\boldsymbol{g}$ designed by an adversary ADV, the performance of the learner is measured by the probability $e_{\mathrm{ADV}(\boldsymbol{g})}(n)$ of not identifying the best arm, i.e., $e_{\mathrm{ADV}(\boldsymbol{g})}(n) \triangleq \mathbb{P}\left(J_n \neq k^\star_{\boldsymbol{g}}\right)$. The smaller $e_{\mathrm{ADV}(\boldsymbol{g})}(n)$, the better the learner. The probability is taken over the randomness of the learner as $\boldsymbol{g}$ is fixed. An alternative definition of the best arm, i.e., with highest G_k in expectation over adversarially sampled $\boldsymbol{g}$, would lead to an impossible problem. Indeed, it could happen that the best arm in expectation is actually always clearly suboptimal on any realization $\boldsymbol{g}$. We define the random variables $\widetilde{\boldsymbol{g}}_{k,t}$ as

$$\widetilde{\boldsymbol{g}}_{k,t} \triangleq \frac{\boldsymbol{g}_{k,t} \mathbf{1}\{I_t = k\}}{\boldsymbol{p}_{k,t}},$$

for arm k at round t and for which $\mathbb{E}_{I_t \sim \boldsymbol{p}_t}\left[\widetilde{\boldsymbol{g}}_{k,t}\right] = \boldsymbol{g}_{k,t}$. We similarly define $\widetilde{G}_{k,t} \triangleq \sum_{t'=1}^{t} \widetilde{\boldsymbol{g}}_{k,t'}$.

Stochastic case In stochastic bandits (Audibert et al., 2010), that we denote STO, the distribution ν_k of arm k is bounded in $[0,1]$ with mean μ_k. The ordering $(\cdot)$ is such that $\mu_{(1)} > \mu_{(2)} \geq \ldots \geq \mu_{(K)}$, as we assume the uniqueness of the best arm without loss of generality. The distributions $\{\nu_k\}$ are unknown to the learner. The best arm to be identified is $k^\star_{\mathrm{STO}} = \mathrm{argmax}_{k \in [K]} \mu_k$. Similar to the adversarial case, the gaps are $\Delta_k \triangleq |\max_{i \neq k} \mu_i - \mu_k|$, ranked as $\Delta_{(1)} \triangleq \Delta_{(2)} \leq \ldots \leq \Delta_{(K)}$, and the error of the learner is $e_{\mathrm{STO}}(n) \triangleq \mathbb{P}\left(J_n \neq k^\star_{\mathrm{STO}}\right)$. However, unlike in the adversarial case, this definition of the error includes the randomness of rewards. Nonetheless, it is only with a probability upper-bounded by $\rho = \mathcal{O}(K \exp(-n\Delta^2_{(1)}))$ that a $\boldsymbol{g} \sim \mathrm{STO}$ can verify $k^\star_{\mathrm{STO}} \neq k^\star_{\boldsymbol{g}}$. However, this difference with the adversarial formulation is not significant as the probability ρ is never larger than the probabilities of errors studied in this paper and it is often insignificant with respect to it.

Notions of complexity Given gaps Δ_k for $k \in [K]$, we define two notions of complexity of the identification task, H_{SR} and H_{UNIF}, that were introduced by Audibert et al. (2010). In particular,

$$H_{\text{SR}} \triangleq \max_{k \in [K]} \frac{k}{\Delta^2_{(k)}} \qquad \text{and} \qquad H_{\text{UNIF}} \triangleq \frac{K}{\Delta^2_{(1)}}.$$

H_{SR} relates to the complexity of the stochastic case. H_{UNIF} will be used both in the adversarial and stochastic cases. In the adversarial case, the complexity and the gaps are defined with respect to $\boldsymbol{g}$ and but we also sometimes write the uniform complexity as $H_{\text{UNIF}}(\boldsymbol{g})$ for clarity.

Class of problems We define a set of classes to group problems with very similar gap structure and with complexities that are only a constant multiplicative factor apart. For any $0 < \Delta_1 = \Delta_2 \leq \Delta_3 \leq \ldots \leq \Delta_K \leq 1/8$, we define a *problem class* $\boldsymbol{\Delta}_c$ with $c \geq 1$. Given these gaps and c, in the adversarial case, we say that $\boldsymbol{g} \in \boldsymbol{\Delta}_c$ if for all $k \in [K]$, $\Delta_k/c \leq \Delta^{\boldsymbol{g}}_k \leq c\Delta_k$ except for only one arm $\overline{k}$ whose gap is related to the smallest gap as $\Delta_1/c \leq \Delta^{\boldsymbol{g}}_{\overline{k}} \leq c\Delta_1$. In the stochastic case, $\text{STO} \in \boldsymbol{\Delta}_c$ under the same condition on its gaps defined as $\Delta_k \triangleq |\max_{i \neq k} \mu_i - \mu_k|$ for $k \in [K]$.

3. General adversarial case: An optimal learner can play uniformly at random

In this section, we define a simple learner (`Rule`) and in Theorem 1, we provide an upper bound on its probability of error depending on the gap in hindsight. We also give a matching lower bound for the general adversarial best-arm identification (Theorem 2) proving that `Rule` obtains the optimal gap-dependent rates of error against worst-case adversaries.

At round t, the *random uniform learner* (`Rule`) selects an arm $I_t \in [K]$ uniformly at random, i.e., with probability $\boldsymbol{p}_{k,t} \triangleq \mathbb{P}(I_t = k) \triangleq 1/K$ for all $k \in [K]$. At the end of the game, the recommended arm J_n is the one with highest estimated cumulative gain, $J_n \triangleq \text{argmax}_{k \in [K]} \widetilde{G}_{k,n}$.

Theorem 1 (Upper bound for `Rule` in the adversarial case) *For any horizon n, given rewards $\boldsymbol{g}$ chosen by an oblivious adversary, with $\boldsymbol{g}_{k,t} \in [0,1]$ for all $t \in [n]$ and for all $k \in [K]$,* `Rule` *outputs an arm J_n with the guarantee that its probability of error $e_{\text{ADV}(\boldsymbol{g})}(n)$ verifies*

$$e_{\text{ADV}(\boldsymbol{g})}(n) \leq K \exp\left(-\frac{3n}{28 H_{\text{UNIF}}(\boldsymbol{g})}\right).$$

Sketch of the proof *(full proof in Appendix B)*: The deviation of $\widetilde{G}_{k,t}$ from $G_{k,t}$ is bounded by a Bernstein bound which applies since, given $\boldsymbol{g}$ and the fact that $\boldsymbol{p}_t$ is fixed to constant values for all the rounds, the $\widetilde{\boldsymbol{g}}_{k,t}$ are independent. $\widetilde{\boldsymbol{g}}_{k,t}$ are scaled Bernoulli random variables where the use of the Bernstein bound leads to a bound that scales with the variance of the $\widetilde{\boldsymbol{g}}_{k,t}$ which is K. Using a Hoeffding bound would lead to a bound that scales with the range of the $\widetilde{\boldsymbol{g}}_{k,t}$ *squared* which is K^2.

Theorem 2 (Lower bound for the adversarial problem) *Given any problem class $\boldsymbol{\Delta}_3$ with associated complexity H_{UNIF}, for any learner, for any horizon n such that $K \exp\left(-n\Delta_1^2/128\right) \leq 1/128$ and $K \geq 4096$, there exist $\boldsymbol{g}^1 \in \boldsymbol{\Delta}_3$ and $\boldsymbol{g}^2 \in \boldsymbol{\Delta}_3$ so that the probabilities of error suffered by the learner on $\boldsymbol{g}^1$ and $\boldsymbol{g}^2$, denoted $e_{\boldsymbol{g}^1}(n)$ and $e_{\boldsymbol{g}^2}(n)$ respectively, verify*

$$\max\left(e_{\boldsymbol{g}^1}(n), e_{\boldsymbol{g}^2}(n)\right) \geq \min\left(\frac{1}{128}\exp\left(-\frac{32n}{H_{\text{UNIF}}}\right), \frac{1}{32}\right).$$

Sketch of the proof *(full proof given in Section D)*: We construct two similar bandit problems. Between the two problems, only one arm differs by a change in the mean of order $\Delta_{(1)}$ for about $n/(2K)$ time steps. Therefore, using a change-of-measure argument, with a probability of order $\exp(-n/H_{\text{UNIF}})$ these two problems generate each a set of rewards, $\boldsymbol{g}^1$ and $\boldsymbol{g}^2$ respectively, that the learner is not able to discriminate. In this undecidable case, the learner still needs to recommend an estimated best arm. However, these two problems have different best arms $k^\star_{\boldsymbol{g}^1} \neq k^\star_{\boldsymbol{g}^2}$. Therefore, the learner makes a mistake of order $\exp(-n/H_{\text{UNIF}})$ on at least one of the two problems.

We consider any *fixed* learner and let us have K base Bernoulli distributions with means $\mu_1 \triangleq 1/2$ and for all $k \in [2:K], \mu_k = 1/2 - \Delta_k$. We consider the first half of the game from round $t = 1$ to a round $\lfloor n/2 \rfloor$ as a set of rounds denoted L. By Dirichlet's box principle, there exists at least one arm, denoted $\overline{k}$, that is pulled less than $\mathcal{O}(n/(2K))$ in expectation during L. This arm, that the learner does not explore very much, is then used to construct the two bandit problems that look similar to the learner. We now describe the two problems in detail.

The first problem is following the original Bernoulli distributions for all arms in phase L. Then the second part of the game, $t = \lfloor n/2 \rfloor + 1, \ldots, n$, is deterministic. Almost all the rewards of all the arms are 0, except some rewards of $\overline{k}$ which are set to 1. This is done so that in expectation in this setup, the total reward of $\overline{k}$ is $n(1/2 - \Delta_1)$ and therefore it becomes the second best arm.

The second problem only differs from the previous one in its first, stochastic part and only affects $\overline{k}$ which instead of having the Bernoulli mean $1/2 - \Delta_{\overline{k}}$, has now a mean of $1/2 - \Delta_{\overline{k}} + 2\Delta_1$. In this problem, the effect of the deterministic part is that in the end, when $t = n$, the expected mean of arm $\overline{k}$ is $(1/2 + \Delta_1)$ instead of $(1/2 - \Delta_1)$, therefore in this case, $\overline{k}$ becomes the best arm.

Remark 3 *The assumption $K \exp(-\Delta^2_{(1)} n/8) \leq 1/32$ is mild. Essentially, it asks for horizon n to be large enough so that the stochastic problem is learnable within n rounds. The assumption on K is likely to be an artifact of the proof. Even with this assumption, our main message holds, in general, no learner can perform better than the random uniform leaner, up to constants.*

4. The best of both worlds challenge

In this section, we ask if we can have a learner that performs optimally under adversarial and stochastic rewards. The lower bound in Theorem 4 shows that in general, this is impossible.

Existing robust solutions? In the stochastic setting, a state-of-the-art algorithm, Sequential Halving (SH, Karnin et al., 2013)—see also Successive Rejects (SR) by Audibert et al. (2010))—guarantees $e_{\text{STO}}(n) \leq \mathcal{O}\left(\log K \exp\left(-n/(H_{\text{SR}} \log K)\right)\right)$. However, as discussed in the introduction, SR or SH can fail against a worst-case adversary. On the other hand, as discussed by Audibert et al. (2010), uniform-like algorithms (like Rule) can only guarantee that in the stochastic case, we get $e_{\text{STO}}(n) \leq \widetilde{\mathcal{O}}\left(\exp\left(-n/H_{\text{UNIF}}\right)\right)$. In general, $H_{\text{SR}} \leq H_{\text{UNIF}}$ and in some problems, we even have $H_{\text{SR}} = H_{\text{UNIF}}/K$. Therefore, SH can notably outperform uniform algorithms in the stochastic case.

The best of both worlds We now reveal the holy grail of our endeavor, which is the following question: Does there exist a learner, unaware of the nature of the reward-generating process, that guarantees for any n, for any stochastic problem STO, and any set of rewards $\boldsymbol{g}$ that its respective probabilities of misidentification $e_{\text{STO}}(n)$ and $e_{\text{ADV}(\boldsymbol{g})}(n)$ simultaneously verify

$$e_{\text{STO}}(n) \leq \widetilde{\mathcal{O}}\left(\exp\left(-\frac{n}{H_{\text{SR}} \log K}\right)\right) \quad \text{and} \quad e_{\text{ADV}(\boldsymbol{g})}(n) \leq \widetilde{\mathcal{O}}\left(\exp\left(-\frac{n}{H_{\text{UNIF}}(\boldsymbol{g})}\right)\right) \text{?}$$

Why is the BOB question challenging? The learner could choose, for arm k, at round t, to use the cumulative gain estimator $\widehat{G}_{k,t} = \frac{t\sum_{t'=1}^{t}\mathbf{1}\{I_{t'}=k\}\boldsymbol{g}_{k,t'}}{\sum_{t'=1}^{t}\mathbf{1}\{I_{t'}=k\}}$. This estimator can be potentially highly biased if it is used against a malignant adversary. For this reason, we base our approach on the estimator $\widetilde{G}_{k,t} = \sum_{t'=1}^{t}\frac{\boldsymbol{g}_{k,t'}}{\boldsymbol{p}_{k,t'}}\mathbf{1}\{I_{t'} = k\}$. However, this usage potentially introduces high variance in our estimates; the final amount of variance of $\widetilde{G}_{k,n}$ is the sum of the variance of each $\widetilde{\boldsymbol{g}}_{k,t}$ and therefore scales with $\sum_{t'=1}^{t} 1/\boldsymbol{p}_{k,t'}$. The high variance is most damaging in the stochastic case when trying to have a learner based on $\widetilde{G}_{k,t}$ to obtain the optimal error rates of Karnin et al. (2013). Indeed, these optimal rates are obtained by algorithms using $\widehat{G}_{k,t}$, which has no bias and small variance in the stochastic case. Therefore, we strive to characterize the minimum amount of unavoidable variance of the mean estimators of each arm. The learner would like to allocate more pulls at any round t to the arms that are among the best arms, which means having large $\boldsymbol{p}_{k,t}$ for these arms. Indeed, discriminating between them is the hardest part of the task and large $\boldsymbol{p}_{k,t}$ reduces the variance term $1/\boldsymbol{p}_{k,t}$. However, it is natural to think that the learner is not able to guarantee that it pulls the best arms at the beginning of the game more than in a uniform fashion. If the arms are pulled uniformly, the variance is of order K, which is very large. The amount of time that the learner accumulates large variance on its estimate of the best arms because they are not yet well identified determines the final probability of error. Intuitively, the lower bound in Theorem 4 constructs worst-case examples showing that any learner cannot pull the best arm more than a certain amount in some period of the game because it is difficult to identify the best arms. Therefore, this learner is susceptible to be tricked by an adversary. Our new learner in Section 5 tries to limit this effect by making early and almost costless bets on what are the best arms given the early rewards and starts to pull them more right away. Note that if any learner allocates pulls uniformly for the first half of as it is done in algorithms like SR or SH, then even if the pulls are randomized, the variance of the estimator $\widehat{G}_k$ of arm k would still scale with K which prevents outperforming even the static random uniform learner.

Another approach could be a learner that determines online if the observed rewards are stochastic. This was used by Bubeck and Slivkins (2012) and Auer and Chiang (2016). They detect if the difference between $\widehat{G}_{k,t}$ and $\widetilde{G}_{k,t}$ is way too large. However, their bound itself uses terms depending on K and the variance of each arm, $1/\boldsymbol{p}_{k,t}$, which leads to similar open questions as discussed just above. In this paper, inspired by the approach of Seldin and Slivkins (2014), we give a practical simple parameter-free and versatile algorithm. Furthermore, the algorithms that are based on stochastic tests are usually cumbersome and complex, as discussed by Seldin and Slivkins, 2014.

Why is the best of both worlds unachievable? We define a new notion of complexity, H_{BOB} as

$$H_{\text{BOB}} \triangleq \frac{1}{\Delta_{(1)}}\max_{k\in[K]}\frac{k}{\Delta_{(k)}}.$$

H_{BOB} is a complexity for the stochastic case. As we detail in Remark 5, $H_{\text{SR}} \leq H_{\text{BOB}} \leq H_{\text{UNIF}}$.

Theorem 4 (Lower bound for the BOB challenge) *For any class problem $\boldsymbol{\Delta}_4$, for any learner, there exists an i.i.d. stochastic problem* STO *with complexity H_{BOB} and there exists an adversarial problem $\boldsymbol{g}$ such that for any n satisfying $K\exp\left(-\Delta_1^2 n/32\right) \leq 1/32$, if the probability of error of the learner on* STO *satisfies*

$$e_{\text{STO}}(n) \leq \frac{1}{64}\exp\left(-\frac{2048n}{H_{\text{BOB}}}\right),$$

then, in the adversarial problem, the learner suffers a constant error,

$$e_{\mathrm{ADV}(\boldsymbol{g})}(n) \geq \frac{1}{16}.$$

Remark 5 *In general, we have*

$$H_{\mathtt{SR}} \leq H_{\mathrm{BOB}} \leq H_{\mathrm{UNIF}}.$$

Below, we compare the three complexities in three specific gap regimes in order to intuitively explore whether the inequalities in the previous equation are strict or not. Interestingly, while in two regimes $H_{\mathtt{SR}} = H_{\mathrm{BOB}}$, in the third regime, called the 'square-root gaps', we can obtain $H_{\mathrm{BOB}} = \sqrt{K/2}H_{\mathtt{SR}}$. This equality shows that on some problems and for large values of K, our lower bound on the complexity of the BOB problem is significantly larger than the complexity of the strictly stochastic case. This ultimately shows that no learner can guarantee the BOB in general and that any learner that is optimal in all strict stochastic problems is then inconsistent against worst-case adversaries.

▶ **Flat regime** We assume all the gaps are equal, $k \in [K]$, $\Delta_k = \Delta_1$. Then, $H_{\mathtt{SR}} = H_{\mathrm{BOB}} = H_{\mathrm{UNIF}}$. Having $H_{\mathtt{SR}} = H_{\mathrm{BOB}}$ shows that our stochastic BOB lower bound for robust learners (Theorem 4) using H_{BOB} is of the same order as the one in the strict stochastic setting (Audibert et al., 2010) using $H_{\mathtt{SR}}$. In this stochastic regime, `Rule` is optimal while being robust to an adversary.

▶ **Super-linear gaps** Let $(2) \in \operatorname{argmin}_{k\in[K]}(\Delta_{(k)}/k)$. This holds if $\forall k \in [3:K]$, we have that $\Delta_{(k)} = k\Delta_{(1)}$, $\Delta_{(1)} \leq 1/K$. Then, $H_{\mathtt{SR}} = H_{\mathrm{BOB}} = (2/K)H_{\mathrm{UNIF}}$. Again, our BOB lower bound is of the same order as in the strict stochastic setting. This seems to indicate that in this case, BOB is achievable. However, it is not achieved by the uniform learner that is clearly suboptimal. This observation demands a new robust learner. Intuitively, the learner can identify bad arms quickly and start focusing early on the best arms without incurring high variance on its estimates for them.

▶ **Square-root gaps** We assume $(2) \in \operatorname{argmin}_{k\in[K]}(\Delta_{(k)}^2/k)$. Let us denote arm j for which $j \in \operatorname{argmin}_{k\in[K]}(\Delta_{(k)}/k)$. For some constant c, let $\Delta_{(1)} = c\Delta_{(j)}/j$. We have $c \leq \sqrt{2j}$ because $\Delta_{(1)}^2/2 \leq \Delta_{(j)}^2/j$ as $(2) \in \operatorname{argmin}_{k\in[K]}(\Delta_{(k)}^2/k)$. Therefore,

$$H_{\mathtt{SR}} = \frac{2}{\Delta_{(2)}^2} = \frac{2}{\Delta_{(1)}^2} = \frac{2}{\Delta_{(1)}}\frac{j}{c\Delta_{(j)}} = \frac{2H_{\mathrm{BOB}}}{c} = \frac{2H_{\mathrm{UNIF}}}{K}.$$

We can get $c = \sqrt{2j} = \sqrt{2K}$. This happens if $\sqrt{K/2}\Delta_{(1)} = \Delta_{(K)}$ and $\sqrt{k/2}\Delta_{(1)} \geq \Delta_{(k)}$ for $k \in [3:K-1]$. Then, we get $\sqrt{K/2}H_{\mathtt{SR}} = H_{\mathrm{BOB}}$. H_{BOB} is $\sqrt{K/2}$ larger than the complexity of the strictly stochastic setting. Intuitively, the learner needs to spend some time to identify the 'square-root gaps' suboptimal arms before starting to focus on the best arms. This makes it suffer an additional amount of variance on its estimates for the best arms.

Sketch of the proof *(full proof in Appendix E)*: Our proof of the lower bound uses some arguments of purely stochastic best-arm identification lower bounds of Audibert et al. (2010) and Carpentier and Locatelli (2016). We have been also inspired by the lower bound of Auer and Chiang (2016) for the BOB question for the cumulative regret. However, our specific construction is new.

Consider a fixed learner. We construct a stochastic and an adversarial problem. Between the two problems, only one arm differs. We bound the number of pulls from the learner on this arm.

Using a change-of-measure argument, with a probability $\mathcal{O}\left(\exp\left(-n/H_{\text{BOB}}\right)\right)$ the two problems are impossible to discriminate. However, the two problems have different best arms. Therefore the learner makes a mistake $\mathcal{O}\left(\exp\left(-n/H_{\text{BOB}}\right)\right)$ on at least one of the two problems.

Let us define K base Bernoulli distributions with means $\mu_1 \triangleq 1/2$ and for all remaining $k \in [2:K]$, $\mu_k \triangleq 1/2 - \Delta_k$. Let $i \in [2:K]$ be an arbitrary arm. Let $n_i \triangleq na_i$, where $a_i \triangleq \Delta_1/\Delta_i \leq 1$, is a number of early rounds in the game. Because the total number of pulls by the learner is limited by n_i during this phase, by Dirichlet's box principle, there exists at least $K - i + 1$ arms included in $[2:K]$ that are pulled by the learner less than $\mathcal{O}(n_i/i)$ times in expectation. Therefore, in this set of arms, there is an arm, denoted $\overline{k}$, that has a gap of order or smaller than Δ_i. This arm, with a small gap, that the learner does not explore very much, is then used to construct the two similar bandit problems that the learner has a hard time to differentiate. The stochastic problem is made by only modifying the original Bernoulli distribution of arm $\overline{k}$ by setting it to $\mu_{\overline{k}} \triangleq 1/2 + \Delta_1/2$. The adversarial problem samples $\boldsymbol{g}$ randomly: It is mimicking the stochastic problem for all rounds and all arms with the exception of the $\overline{k}$ during the first n_i rounds of the game where the gains are from the base Bernoulli distribution (with mean $1 - \Delta_{\overline{k}}$).

If $a_i \geq \Delta_1/\Delta_i$— fixing a large enough phase at the beginning to modify the identity of the best arm—then the best arms in both problems are different. The event on which the two problems are impossible to discriminate has a probability of $\mathcal{O}\left(\exp(-n_i(\Delta_i)^2/i\right)$. To maximize this probability, we minimize a_i while still ensuring to have the change of best arm between the two problems, by setting $a_i \triangleq \Delta_1/\Delta_i$. Therefore, the probability is now $\mathcal{O}\left(\exp(-n\Delta_1\Delta_i/i)\right)$. Again to maximize it, we choose $i \triangleq \operatorname{argmin}_k(\Delta_k/k)$ and obtain the claimed result for some fixed $\boldsymbol{g}$.

5. A simple robust parameter-free algorithm for stochastic & adversarial rewards

In this section, we present a new learner and analyze its theoretical performance against any i.i.d. stochastic problem or any adversarially designed rewards.

We call the algorithm `ProbabilityONE`, denote it by `P1`, and detail it in Figure 2. Intuitively, `P1` pulls the estimated best arm with "probability" one, the estimated second best arm with "probability" one half, the estimated third best arm with "probability" one third, and so on until pulling the estimated worst arm with "probability" one over K.[a] In order to get proper probabilities, we need to normalize them by the K-th harmonic number $\overline{\log} K = \sum_{k=1}^{K}(1/k)$, where $\overline{\log} K \leq \log K + 1$ for all positive integers K. `P1` is following a Zipf distribution with an exponent of 1 (Powers, 1998).

For $t = 1, 2, \ldots$

- Sort and rank the arms by decreasing order of $\widetilde{G}_{\cdot,t-1}$: Rank arm k as $\widetilde{\langle k \rangle}_t \in [K]$[a].
- Select arm $I_t \in [K]$ with $\boldsymbol{p}_{k,t} \triangleq \mathbb{P}\left(I_t = k\right) \triangleq \dfrac{1}{\widetilde{\langle k \rangle}_t \, \overline{\log} K}$ for all $k \in [K]$.

Recommend, at any given round t, $J_t \triangleq \operatorname{argmax}_{k \in [K]} \widetilde{G}_{k,t}$.

a. Equalities between arms or comparisons with arms that have not been pulled yet are broken arbitrarily.

Figure 2: The `ProbabilityONE` (`P1`) algorithm

The estimate used in P1 is $\widetilde{G}_{k,t-1}$ for arm k at round t. P1 is heavily inspired by Successive Rejects (SR) of Audibert et al. (2010), as both are somehow ranking the arms and attempt to allocate to arm k a number of pulls $\widetilde{\mathcal{O}}\left(n/\langle k\rangle\right)$ according to its true rank $\langle k\rangle$. Our new learner is parameter-free and naturally anytime (agnostic of n). As SR, it does not need any knowledge of any complexity nor it is trying to estimate any. However, contrarily to SR, it does not divide the game into different sampling phases. The same rudimentary sampling procedure is repeated at all rounds t in æternum.

As discussed in Section 4, in order to minimize the misidentification error in the stochastic case, it is crucial to limit the variance of the estimators for the best arms. Therefore P1, from its very first pull, pulls more (i.e., with higher probability) the arms that are estimated to be among the best. First, this comes with almost no cost: Indeed, pulling the estimated best arm with probability $1/\overline{\log}K$ does not prevent from pulling all the arms almost uniformly and more precisely with probability at least $1/(K\overline{\log}K)$. Therefore, no suboptimal arm is actually left out in the early chase for the best arm and the variances of the estimators can only increase by a factor of $\overline{\log}K$ with respect to the uniform strategy. Second, it gives the learner the flexibility to adapt to different types of stochastic problems with different gap regimes. If in a setup, some arms are clearly suboptimal, it is helpful to pull the clearly best arm more and right from the beginning. This is a more flexible behavior than the one of algorithms that are fixing the number of pulls for each arm during a fixed period in advance. Additionally, compared to a fixed-phase algorithm, our analysis is also more flexible: We can analyze, for instance, the quality of the estimated ranking $\widetilde{\langle\cdot\rangle}$ and therefore the adaptive sampling procedure of the arms at any round. Actually, these rounds called *comparison rounds*, can be chosen in a problem-dependent manner, in order to minimize the final probability of error. This is conspicuous in the complexity measure present in the upper bound as it is defined as a minimum complexity among complexities defined for any set of comparison rounds. Note again that this optimization procedure is only for the analysis of the learner while the learner itself is utterly agnostic of the optimal 'virtual' phases and just follows its simple routine. We now define this new notion of complexity. First, we define the proportion of rounds for comparison $\boldsymbol{a}$ in a space $\boldsymbol{A}$. Let $\boldsymbol{A} \triangleq \left\{\boldsymbol{a} \in [0,1]^K : na_i \in \mathbb{N}, \forall i \in [K], 1 = a_1 = a_2 \geq a_3 \geq \ldots \geq a_K > a_{K+1} = 0\right\}$. The complexity associated with the P1 learner is H_{P1} and is defined first as

$$H_{\text{P1}}(\boldsymbol{a}) \triangleq \max_{k\in[K]} \frac{\sum_{i=\langle k\rangle}^{K}(a_i - a_{i+1})i + \frac{1}{24}Ka_{\langle k\rangle}\Delta_k}{a_{\langle k\rangle}^2\Delta_k^2}\overline{\log}K \quad \text{and then,} \quad H_{\text{P1}} \triangleq \min_{\boldsymbol{a}\in\boldsymbol{A}} H_{\text{P1}}(\boldsymbol{a}).$$

In $H_{\text{P1}}(\boldsymbol{a})$, for arm k, the term $\sum_{i=\langle k\rangle}^{K}(a_i - a_{i+1})i$ corresponds to the sum of variances $\widetilde{\mathcal{O}}(i)$, for $i \in [\langle k\rangle : K]$, during a proportion of time $a_i - a_{i+1}$ between the comparison rounds $i+1$ and i. Indeed, we expect the estimated ranking of arm k, $\widetilde{\langle k\rangle}_t$, for $t \in [na_{i+1} : na_i]$, to be smaller than i, which corresponds to its true ranking as $\langle k\rangle \leq i$. This leads, for $t \geq na_{i+1}$, to $\boldsymbol{p}_{k,t} \geq 1/(i\overline{\log}K)$ and therefore, as $\widetilde{\boldsymbol{g}}_{k,t} = \frac{\boldsymbol{g}_{k,t}}{\boldsymbol{p}_{k,t}}\mathbf{1}\{I_t = k\}$, to a variance of $\widetilde{\boldsymbol{g}}_{k,t}$ smaller than $i\overline{\log}K$. In the denominator, the term $a_{\langle k\rangle}$ is proportional to the amount of pulls, $a_{\langle k\rangle}n$, allocated to arm k.

Theorem 6 (Upper bounds for P1) *For any stochastic problem* STO *with complexity H_{P1} and for any $\boldsymbol{g}$ fixed by an oblivious adversary with complexity $H_{\text{UNIF}}(\boldsymbol{g})$, the probabilities of error of* P1*, denoted $e_{\text{STO}}(n)$ and $e_{\text{ADV}(\boldsymbol{g})}(n)$ in their respective environment, for any n, simultaneously verify*

$$e_{\text{STO}}(n) \leq 2K^3 n \exp\left(-\frac{n}{128H_{\text{P1}}}\right) \quad \textit{and} \quad e_{\text{ADV}(\boldsymbol{g})}(n) \leq K\exp\left(-\frac{3n}{40\overline{\log}(K)H_{\text{UNIF}}(\boldsymbol{g})}\right).$$

Sketch of the proof *(full proof in Appendix F)*: For the *adversarial case*, it is enough that the learner pulls each arm with a probability larger than $1/(K\overline{\log}K)$ to obtain the same complexity H_{UNIF} as `Rule`, up to a factor $\log K$. For the *stochastic case*, we consider $K-1$ arbitrary 'virtual' phases that each ends at round $n_i = na_i$, that will be chosen in hindsight to minimize the upper bound. Note that `P1` is oblivious to these values. The phases are following a countdown from phase K to phase 2 that is the last one. Intuitively, n_i is a round after which we expect the following event ξ_i to happen with high probability: For all $t > n_i$, `P1` has well estimated the rank of any arm k with a significantly smaller gap than the i-th gap, in particular, $\widetilde{\langle k \rangle}_t \leq i-1 < i$, if $\mu_{(1)} - \mu_k \leq \Delta_{(i)}/2$. The important consequence is that any such arm k, for $t > n_i$, will be pulled with $\boldsymbol{p}_{k,t} \geq 1/(i-1)$ leading to a smaller variance (of order $i-1$) in their estimates $\widetilde{\boldsymbol{g}}_{k,t}$ for $t > n_i$. Reducing these variances leads in turn to better estimates in the rest of the game. The proof works iteratively over the phases. We consider that an error has occurred as soon as the estimated ranking is wrong at the end of a phase i, i.e., that ξ_i does not hold. We bound the probability of making such a mistake at the end of phase i, give the fact that no mistakes were made in previous phases. Indeed, with no past mistake before phase i, the learner is guaranteed to have sharp estimates. Summing all the errors gives a bound on the probability of not ranking well the best arm at the end of the last phase $n_2 = n$.

To bound $e_{\text{STO}}(n)$, we use the Bernstein inequality for martingale differences. This inequality takes into account the variances and holds despite the dependencies of the random variables $\widetilde{\boldsymbol{g}}_{k,t}$.

When minimizing $H_{\texttt{P1}}$ by choosing a_k, for $k \in [K]$, we need to trade off between short phases, possibly meaning not enough samples to discriminate the suboptimal arms (the denominator term of the suboptimal arms is small) and long phases, which means that in early stages, the best arms are considered as badly ranked for a long time and the variance of their mean estimators is increasing with the length of the early phases (the numerator term for the best arms is larger).

Corollary 7 *The complexity $H_{\texttt{P1}}$ of `P1` matches the complexity H_{BOB} from the lower bound of Theorem 4 of up to log factors,*

$$H_{\texttt{P1}} = \mathcal{O}\left(H_{\text{BOB}} \log^2 K\right).$$

The result of Corollary 7 is obtained by setting $a_k = \Delta_{(1)}/\Delta_{(k)}$, $\forall k \in [K]$.[3] Notice that the same values were also used in the lower bound in Theorem 4. The full proof of Corollary 7 is in Appendix G, where we also discuss the relation between $H_{\texttt{P1}}$ and $H_{\texttt{SR}}$ and H_{BOB} for different regimes of the gaps. Corollary 7 demonstrates that `P1` achieves the best that can be wished for in the two worlds, up to log factors.

Remark 8 *In the adversarial case, a modification to `P1` leads to similar upper bound as for `Rule`, where $H_{\text{UNIF}}(\boldsymbol{g})$ appears instead of $H_{\text{UNIF}}(\boldsymbol{g})\overline{\log}K$. Indeed, with probability one half we can play according to `Rule` and otherwise use `P1`. We keep the recommendation $J_n = \operatorname{argmax}_{k\in[K]} \widetilde{G}_{k,n}$.*

Remark 9 *We studied the hard (adversarial) and easy data (stochastic) settings. However, as discussed by Seldin and Slivkins (2014) and Allesiardo et al. (2017), we can consider intermediate settings of difficulty. First, quite simply, the result in the stochastic case would still hold up to constants when the gaps of the arms do not change by more than the same numerical factor during the game. More generally, we could design variants of $H_{\texttt{P1}}$ under the assumption that after some round t' the (ground-truth) ranks of all the arms are upper bounded each by a constant. Indeed, soon after t', `P1` will itself rank every arm at most according to its maximum rank. Such results would even hold in a case of a change of the identity of the best arm in the game.*

3. To ease the exposition, we assume without loss of generality that $na_i \in \mathbb{N}$, $\forall i \in [K]$.

6. The simplicity of P1 and its sampling routine have potential extensions

Fixed-confidence In this i.i.d. stochastic setting, (Maron and Moore, 1993, Even-Dar et al., 2006; Mnih et al., 2008; Kalyanakrishnan et al., 2012; Kaufmann and Kalyanakrishnan, 2013; Garivier and Kaufmann, 2016) the goal is to design a learner that stops as soon as possible and returns the best arm with a fixed confidence. Let $\widetilde{n}$ be the round when the algorithm stops and $J_{\widetilde{n}}$ its returned arm. Given a confidence level δ, the learner has to guarantee that $\mathbb{P}\left(J_{\widetilde{n}} = k^\star_{\text{STO}}\right) \leq \delta$. The performance of the learner is then measured by its *sample complexity*, which is the number of rounds $\widetilde{n}$ before stopping, either in expectation or in high probability.

Mimicking the Hoeffding and Bernstein Races (Maron and Moore, 1993; Mnih et al., 2008), we could design Freedman Races based on P1. All the arms would be pulled according to the parameter-free sampling routine of P1. No arm could be ever discarded because we could happen to face an adversary. The learner would stop using the P1 routine based on having the confidence intervals from the Bernstein concentration inequality for martingales and in particular, it would stop when the confidence interval for the empirically best arm is separated from the confidence intervals for all the other arms. For this fixed-confidence variant of P1, we could reuse the proof techniques developed for the fixed-budget setting and bound the accumulation of variance of the estimates. Then, in the stochastic case, we would be able to guarantee that the expected sample complexity of such algorithm is $\widetilde{\mathcal{O}}\left(H_{\text{BOB}} \log(1/\delta)\right)$, up to log factors.

For the adversarial case, we can consider an infinite sequence of rewards $\boldsymbol{g}$ fixed by the adversary for all arms. Assume that the sample complexity on the rewards up to round t, is bounded by some $\boldsymbol{n}(\boldsymbol{g}_{[t]})$, the smaller the better. We can then guarantee that if at any round $\widetilde{n}$, $\boldsymbol{g}_{[\widetilde{n}]}$ verifies $\boldsymbol{n}(\boldsymbol{g}_{[\widetilde{n}]}) \leq \widetilde{n}$, then with probability $1-\delta$, the learner can both stop and recommend the best arm $k^\star_{\boldsymbol{g}_{[\widetilde{n}]}}$ at round $\widetilde{n}$. Our Freedman algorithm would be able to satisfy this requirement with the complexity of uniform allocation, $\widetilde{\mathcal{O}}\left(H_{\text{UNIF}}(\boldsymbol{g}_{[\widetilde{n}]}) \log(1/\delta)\right)$. Note that the learner could possibly never stop.

Streams, windows, thresholds, m-set, and active anomaly detection P1 can be used to recommend the best arm in the latest time window between $t-W$ and t for each round t. Essentially, W would replace n in our bounds if P1 recommends the estimated best arm in that window. Also, P1 could also be adapted to identify m best arms out of K as did Bubeck et al. (2013) The key is to redefine the gap with respect to the m-th best arm instead of the best arm. We could also extend P1 to a setting where the rewards converge (Li et al., 2016). Locatelli et al. (2016) defined the gaps with respect to a given fixed threshold and the objective is to determine which arms have a mean higher than the threshold. Again, our approach would apply. Moreover, it could prove to be a good robust approach to the problems that are linked to the threshold bandit problem as discussed by Locatelli et al. (2016, Section 3), one of them being the *active anomaly detection* (Carpentier and Valko, 2014). Indeed, in adversarial anomaly detection, the learner might be monitoring different streams of non-stochastic rewards and could potentially detect an anomaly if one of the streams outputs a reward signal that is on average larger than a given threshold during a time window period W. A variant of P1 would then be a robust anomaly detector and would be also adaptive to easy data.

Adaptive adversaries Our upper bounds results extend naturally to adaptive adversaries given the following condition: H_{UNIF} is an upper bound on the complexity of all $\boldsymbol{g}$ that the adaptive adversary can possibly generate. The proofs remain the same except that the Bernstein concentration inequality in Theorem 1 is replaced by the Bernstein concentration inequality for martingales.

Acknowledgements

We gratefully acknowledge the support of the NSF through grant IIS-1619362 and of the Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). The research presented was also supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, Inria and Otto-von-Guericke-Universität Magdeburg associated-team north-european project Allocate, and French National Research Agency projects ExTra-Learn (n.ANR-14-CE24-0010-01) and BoB (n.ANR-16-CE23-0003). We would like to thank Iosif Pinelis for a useful discussion on Bernstein inequalities.

References

Robin Allesiardo and Raphaël Féraud. Selection of learning experts. In *International Joint Conference on Neural Networks*, 2017.

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 2017.

Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best-arm identification for contaminated bandits. *arXiv preprint* `arXiv:1802.09514`, 2018.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best-arm identification in multi-armed bandits. In *Conference on Learning Theory*, 2010.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory and arXiv preprint* `arXiv:1605.08722`, 2016.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1), 2002.

Sébastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, 2013.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1), 2012.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, 2012.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandit problems. In *Algorithmic Learning Theory*, 2009.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best-arm identification bandit problem. In *Conference on Learning Theory*, 2016.

Alexandra Carpentier and Michal Valko. Extreme bandits. In *Neural Information Processing Systems*, 2014.

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*, 7:1079–1105, 2006.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

Aurélien Garivier and Emilie Kaufmann. Optimal best-arm identification with fixed confidence. In *Conference on Learning Theory*, 2016.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best-arm identification and hyperparameter optimization. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Rob Kaas and Jan M. Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1):13–18, 1980.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, 2013.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, 2013.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. `Hyperband:` A novel bandit-based approach to hyperparameter optimization. *arXiv preprint* `arXiv:1603.06560`, 2016.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, 2016.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun), 2004.

Oded Maron and Andrew Moore. Hoeffding Races: Accelerating model-selection search for classification and function approximation. In *Neural Information Processing Systems*, 1993.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *International Conference on Machine Learning*, 2008.

David Powers. Applications and explanations of Zipf's law. In *New methods in language processing and computational natural language learning*. Association for Computational Linguistics, 1998.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the `EXP3++` algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, 2014.

Appendix A. Experiments in the stochastic setting

After the theoretical main course, we propose an experimental dessert. We reuse the experimental setups of Audibert et al. (2010) in Experiments 1 to 7. We only consider Bernoulli distributions and the optimal arm has always mean $1/2$. Each experiment corresponds to a different situation for the gaps. They are either clustered in a few groups or distributed according to an arithmetic or geometric progression. Experiment 8 reuses the 'square-root gap' scenario when $H_{\text{BOB}} = \sqrt{2K}H_{\text{SR}}$ as detailed in Remark 5. The experimental setups are given below.

- Experiment 1: *One group of bad arms*, $K = 20$, $\mu_{2:20} = 0.4 \equiv \forall j \in \{2, \dots, 20\}, \mu_j = 0.4$
- Experiment 2: *Two groups of bad arms*, $K = 20$, $\mu_{2:6} = 0.42$, $\mu_{7:20} = 0.38$.
- Experiment 3: *Geometric progression*, $K = 4$, $\mu_i = 0.5 - (0.37)^i$, $i \in \{2, 3, 4\}$
- Experiment 4: *6 arms divided into three groups*, $K = 6$, $\mu_2 = 0.42$, $\mu_{3:4} = 0.4$, $\mu_{5:6} = 0.35$
- Experiment 5: *Arithmetic progression*, $K = 15$, $\mu_i = 0.5 - 0.025i$, $i \in \{2, \dots, 15\}$
- Experiment 6: *2 good arms and a large group of bad arms*, $K = 20$, $\mu_2 = 0.48$, $\mu_{3:20} = 0.37$
- Experiment 7: *Three groups of bad arms*, $K = 30$, $\mu_{2:6} = 0.45$, $\mu_{7:20} = 0.43$, $\mu_{21:30} = 0.38$
- Experiment 8: *Square-root gaps* $K = 100$, $\mu_i = 0.5 - 0.25\sqrt{i/(2K)}$, $i \in [2 : 100]$

In Table 1, we report the complexities H_{SR}, H_{BOB}, and H_{UNIF} computed in these experimental setups. Unsurprisingly, in Experiments 1, 3, and 5 we recover $H_{\text{SR}} = H_{\text{BOB}}$ and in Experiment 8, we have $H_{\text{BOB}} = \sqrt{2K}H_{\text{SR}}$. Experiments 2, 4, 6, and 7 then give an idea about the behavior of H_{SR}, H_{BOB}, and H_{UNIF} with respect to each other.

Experimental setup	H_{SR}	H_{BOB}	H_{UNIF}
1. One group of bad arms	2000	2000	2000
2. Two groups of bad arms	1389	2083	3125
3. Geometric progression	5540	5540	11080
4. 6 arms divided into three groups	400	500	938
5. Arithmetic progression	3200	3200	24000
6. 2 good arms and a large group of bad arms	5000	7692	50000
7. Three groups of bad arms	4082	5714	12000
8. Square-root gaps	3200	22627	160000

Table 1: Comparing complexities H_{SR}, H_{BOB}, and H_{UNIF}.

In Figure 3, we report the average success rate (which is an estimate of the probabilities of error) of `SR`, `P1`, and the *static uniform allocation* on the 8 experimental problems previously detailed. The static uniform allocation is not the algorithm `Rule`. `Rule` samples an arm uniformly at random while the *static* uniform allocation simply allocates n/K pulls to each arm deterministically. The empirical results follow very closely our theoretical findings as the empirical behavior in Figure 3 mimics the behavior of the complexities in Table 1. As Audibert et al. (2010), we chose horizon n to be of the order of the complexity $H_1 = \sum_{k \in [K]}(1/\Delta_k^2)$, where $H_{\text{SR}} \leq H_1 \leq H_{\text{SR}} \log K$.

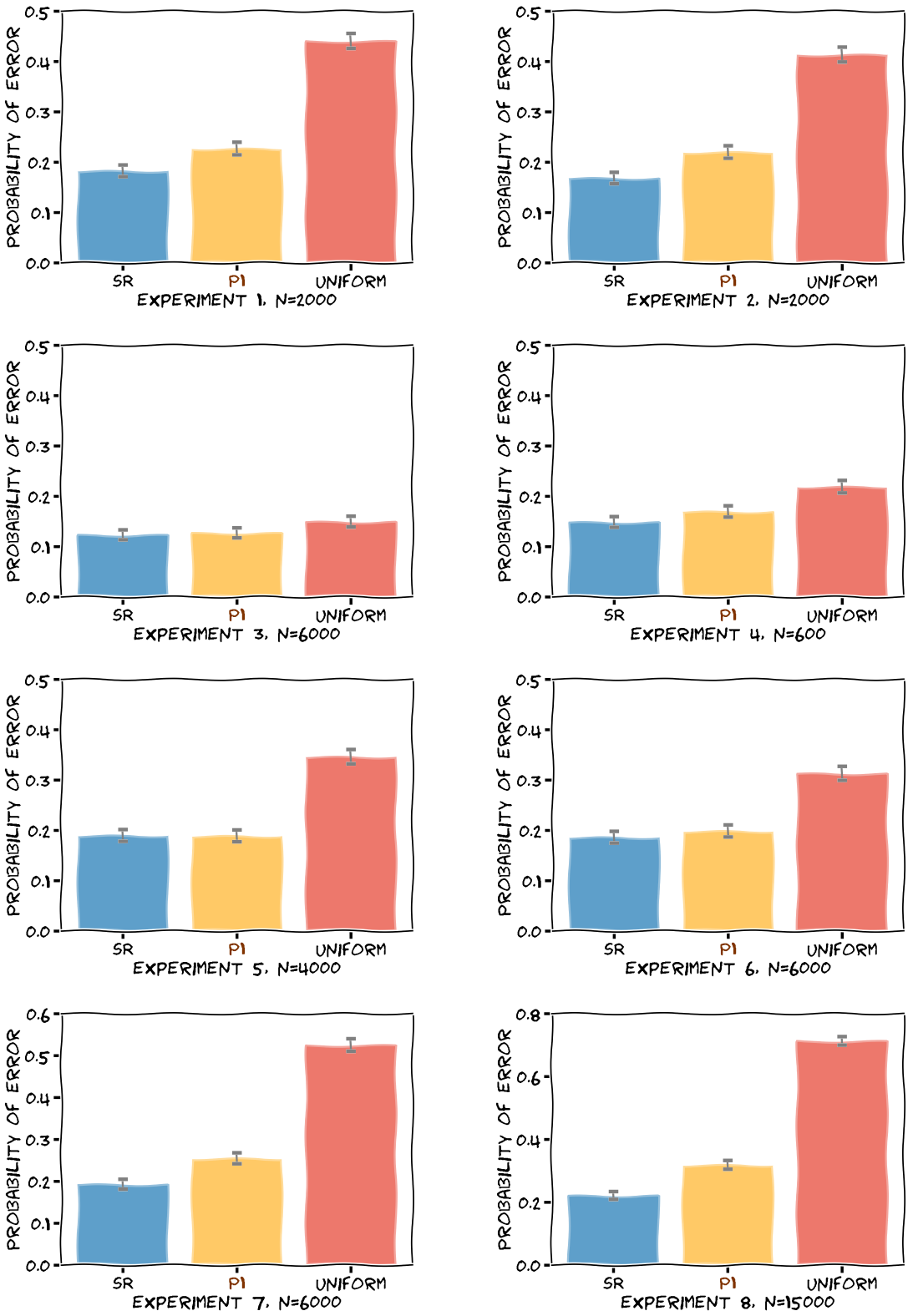

Figure 3: Probabilities of error of SR, P1, and the static uniform allocation.

In the remained of the appendix, for a random variable X, we denote its variance by σ_X^2. Moreover, we also write that a bounded random variable $X \in [\ell_X, u_X]$ has a range $b_X = u_X - \ell_X$.

Appendix B. Upper bound on the probability of error of Rule in the general adversarial case

Theorem 1 (Upper bound for Rule in the adversarial case) *For any horizon n, given rewards $\boldsymbol{g}$ chosen by an oblivious adversary, with $\boldsymbol{g}_{k,t} \in [0,1]$ for all $t \in [n]$ and for all $k \in [K]$,* Rule *outputs an arm J_n with the guarantee that its probability of error $e_{\mathrm{ADV}(\boldsymbol{g})}(n)$ verifies*

$$e_{\mathrm{ADV}(\boldsymbol{g})}(n) \leq K \exp\left(-\frac{3n}{28 H_{\mathrm{UNIF}}(\boldsymbol{g})}\right).$$

Proof We assume that the arms are sorted by their means so that arm 1 is the best. Given the adversary gain vector $\boldsymbol{g}$, the random variables $\widetilde{\boldsymbol{g}}_{k,t}$ are conditionally independent from each other for all $k \in [K]$ and $t \in [n]$ as we have $\boldsymbol{p}_{k,t} \triangleq 1/K$, fixed for all $k \in [K]$ and $t \in [n]$. We have

$$\begin{aligned}
e_{\mathrm{ADV}(\boldsymbol{g})}(n) &\triangleq \mathbb{P}\left(J_n \neq k^\star_{\boldsymbol{g}}\right) = \mathbb{P}\left(\exists k \in [2:K] : \widetilde{G}_{1,n} \geq \widetilde{G}_{k,n} \middle| \boldsymbol{g}\right) \\
&\leq \mathbb{P}\left(\exists k \in [2:K] : \widetilde{G}_{k,n} - G_k \geq \frac{n\Delta_k^{\boldsymbol{g}}}{2} \text{ or } \widetilde{G}_{1,n} - G_1 \leq \frac{n\Delta_1^{\boldsymbol{g}}}{2} \middle| \boldsymbol{g}\right) \\
&\leq \mathbb{P}\left(\widetilde{G}_{1,n} - G_1 \leq \frac{n\Delta_1^{\boldsymbol{g}}}{2} \middle| \boldsymbol{g}\right) + \sum_{k=2}^{K} \mathbb{P}\left(\widetilde{G}_{k,n} - G_k \geq \frac{n\Delta_k^{\boldsymbol{g}}}{2} \middle| \boldsymbol{g}\right) \\
&\overset{\text{(a)}}{\leq} \sum_{k=1}^{K} \exp\left(-\frac{3(\Delta_k^{\boldsymbol{g}})^2 n}{28K}\right) \\
&\leq K \exp\left(-\frac{3(\Delta_1^{\boldsymbol{g}})^2 n}{28K}\right),
\end{aligned}$$

where **(a)** is using Bernstein's inequality applied to the sum of the random variables with mean zero that are $\widetilde{\boldsymbol{g}}_{k,t} - \boldsymbol{g}_{k,t}$ for which we have the following bounds on the variance and range. The variance of $\widetilde{\boldsymbol{g}}_{k,t}$ is the variance of a scaled Bernoulli random variable with parameter $1/K$ and range $[0, K\boldsymbol{g}_{k,t}]$, therefore we have $|\widetilde{\boldsymbol{g}}_{k,t} - \boldsymbol{g}_{k,t}| \leq K$, and $\sigma^2_{\widetilde{\boldsymbol{g}}_{k,t} - \boldsymbol{g}_{k,t}} = \sigma^2_{\widetilde{\boldsymbol{g}}_{k,t}} = 1/K(1-1/K)K^2\boldsymbol{g}_{k,t}^2 \leq K$. We also use $\Delta_k^{\boldsymbol{g}} \leq 1$, for all $k \in [K]$ so that we have for all $k \in [2:K]$ (and a symmetrical argument for $k = 1$),

$$\begin{aligned}
\mathbb{P}\left(\widetilde{G}_{k,n} - \widetilde{G}_{k,n} - n\Delta^{\frac{\boldsymbol{g}_k}{2}} \leq -\frac{n\Delta_k^{\boldsymbol{g}}}{2}\right) &\leq \exp\left(-\frac{(\Delta_k^{\boldsymbol{g}}/2)^2 n^2/2}{\sum_{t=1}^n \sigma^2_{\widetilde{\boldsymbol{g}}_{k,t} - \boldsymbol{g}_{k,t}} + \frac{1}{6}K\Delta_k^{\boldsymbol{g}} n}\right) \\
&\leq \exp\left(-\frac{(\Delta_k^{\boldsymbol{g}})^2 n^2/8}{nK + \frac{1}{6}Kn}\right) \\
&\leq \exp\left(-\frac{(\Delta_k^{\boldsymbol{g}})^2 n^2/8}{nK + \frac{1}{6}Kn}\right) \\
&= \exp\left(-\frac{3(\Delta_k^{\boldsymbol{g}})^2 n}{28K}\right).
\end{aligned}$$

■

Appendix C. Change of measure

Lemma 10 (Change of measure) *Let L be a phase, i.e., a subset of rounds of the game, $L \subset [n]$. Let us consider two bandit problems. In these two problems, at all rounds $t \in [n]$, and for all arms $k \in [K]$, the rewards $\boldsymbol{g}_{k,t}$ are sampled in a stochastic learner-oblivious independent fashion from a distribution $\nu_{k,t}$. We consider problems that for all rounds of the game only differ in the rewards of one arm $\overline{k}$ during phase L. For all $t \in L$, in the two problems, the distribution of arm $\overline{k}$ is a Bernoulli (independent of all the other events in the bandit game for any round $t \in L$) with means $\mu^2_{\overline{k}}(t) \triangleq \mu^2_{\overline{k}} \triangleq 1/2 + \Delta$ and $\mu^1_{\overline{k}}(t) \triangleq \mu^1_{\overline{k}} \triangleq 1/2 - \Delta'$ respectively for the two problems, where $1/8 > \Delta' \geq \Delta \geq 0$. The expectation and probability with respect to the learner and the samples of this problem p are denoted by $\mathbb{E}_p$ and $\mathbb{P}_p$. Then, if we have an event W depending only on $\boldsymbol{g}$ generated by the problems and the actions of the learner $I_{[n]}$ when the number of rounds arm $\overline{k}$ is pulled during phase L is upper-bounded by B, we have*

$$\mathbb{P}_2(W) \geq \frac{\mathbb{P}_1(W)}{8} \exp\left(-16\left(\Delta'\right)^2 B\right).$$

Proof This lemma is a slight extension of Lemma 12 by Auer and Chiang (2016) which in turn is based on the result of Mannor and Tsitsiklis (2004). In the case of Auer and Chiang (2016), $\Delta' \triangleq \Delta$.

For $p \in \{1, 2\}$, let $\mathbb{P}_p$ and $\mathbb{E}_p$ denote the probability and expectation with respect to the bandit problem p defined above. Let $\overline{G} \triangleq \sum_{t \in L} \boldsymbol{g}_{\overline{k},t} \mathbf{1}\{I_t = \overline{k}\}$ be the sum of rewards received when playing arm $\overline{k}$ in phase L. Conditioned on the number of pulls of $\overline{k}$ during phase L that we denote by $\overline{T}$, $\overline{G}$ is a binomial random variable with parameters $\overline{T}$ and $\mu^p_{\overline{k}}$ in problem p. Hence, by Kaas and Buhrman (1980),

$$\mathbb{P}_1\left(\overline{G} \leq \lfloor \overline{T}(1/2 - \Delta') \rfloor\right) \leq \frac{1}{2}.$$

Let w denote a particular realization of rewards $\boldsymbol{g}_{k,t}$, $k \in [K]$, $t \in L$, and learner choices $\{I_t\}_{t \in L}$. For any realization w, the probabilities $\mathbb{P}_1(w)$ and $\mathbb{P}_2(w)$ are related by

$$\mathbb{P}_2(w) = \mathbb{P}_1(w) \frac{(1/2 + \Delta)^{\overline{G}(w)} (1/2 - \Delta)^{\overline{T}(w) - \overline{G}(w)}}{(1/2 - \Delta')^{\overline{G}(w)} (1/2 + \Delta')^{\overline{T}(w) - \overline{G}(w)}}.$$

Also, since $\frac{1/2+\Delta}{1/2-\Delta'} \geq 1$ and both the numerator and denominator of the previous fraction are positive, then the function $x \mapsto \frac{1/2+\Delta+x}{1/2-\Delta'+x}$ is non-increasing for $x \geq 0$. Therefore, we have

$$\frac{1/2 + \Delta}{1/2 - \Delta'} \geq \frac{1/2 + \Delta + (\Delta' - \Delta)/2}{1/2 - \Delta' + (\Delta' - \Delta)/2} = \frac{1/2 + (\Delta' + \Delta)/2}{1/2 - (\Delta' + \Delta)/2}.$$

Similarly, since $\frac{1/2-\Delta}{1/2+\Delta'} \leq 1$ and both the numerator and denominator of the previous fraction are positive, then the function $x \mapsto \frac{1/2-\Delta-x}{1/2+\Delta'-x}$ is non-increasing for $1/2 - \Delta \geq x \geq 0$. Therefore, choosing $x = (\Delta' - \Delta)/2$, which verifies $x = (\Delta' - \Delta)/2 \leq 1/2 - \Delta$ we have that

$$\frac{1/2 - \Delta}{1/2 + \Delta'} \geq \frac{1/2 - \Delta - (\Delta' - \Delta)/2}{1/2 + \Delta' - (\Delta' - \Delta)/2} = \frac{1/2 - (\Delta' + \Delta)/2}{1/2 + (\Delta' + \Delta)/2}.$$

Therefore, we have

$$\mathbb{P}_2(w) \geq \mathbb{P}_1(w)\frac{(1+\Delta+\Delta')^{\overline{G}(w)}(1-\Delta-\Delta')^{\overline{T}(w)-\overline{G}(w)}}{(1-\Delta'-\Delta)^{\overline{G}(w)}(1+\Delta'+\Delta)^{\overline{T}(w)-\overline{G}(w)}}$$
$$= \mathbb{P}_1(w)\left(\frac{1-\Delta-\Delta'}{1+\Delta+\Delta'}\right)^{\overline{T}(w)-2\overline{G}(w)} \geq \mathbb{P}_1(w)\left(\frac{1-2\Delta'}{1+2\Delta'}\right)^{\overline{T}(w)-2\overline{G}(w)}.$$

If $\overline{G}(w) \geq \lfloor \overline{T}(w)(1/2-\Delta') \rfloor$, then $\mathbb{P}_2(w) \geq \mathbb{P}_1(w)\left(\frac{1-2\Delta'}{1+2\Delta'}\right)^{2\Delta'\overline{T}(w)+2}$.
Hence,

$$\begin{aligned}
\mathbb{P}_2(W) &\geq \mathbb{P}_2\left(W \cap \overline{G} \geq \lfloor \overline{T}(1-\Delta') \rfloor\right) \\
&\geq \mathbb{P}_1\left(W \cap \overline{G} \geq \lfloor \overline{T}(1-\Delta') \rfloor\right)\left(\frac{1-2\Delta'}{1+2\Delta'}\right)^{2\Delta' B+2} \\
&\geq \frac{\mathbb{P}_1(W)}{2}\left(\frac{1-2\Delta'}{1+2\Delta'}\right)^{2\Delta' B+2} \\
&\geq \frac{\mathbb{P}_1(W)}{2}\left(1-4\Delta'\right)^{2\Delta' B+2} \\
&\overset{\text{(a)}}{\geq} \frac{\mathbb{P}_1(W)}{8}\exp\left(-16(\Delta')^2 B\right),
\end{aligned}$$

where **(a)** is because for $0 \leq x \leq 1/2$, $1-x \geq e^{-2x}$ and $\Delta' \leq 1/8$. ■

Appendix D. Lower bound in the general adversarial setting

Theorem 2 (Lower bound for the adversarial problem) *Given any problem class $\boldsymbol{\Delta}_3$ with associated complexity H_{UNIF}, for any learner, for any horizon n such that $K\exp\left(-n\Delta_1^2/128\right) \leq 1/128$ and $K \geq 4096$, there exist $\boldsymbol{g}^1 \in \boldsymbol{\Delta}_3$ and $\boldsymbol{g}^2 \in \boldsymbol{\Delta}_3$ so that the probabilities of error suffered by the learner on $\boldsymbol{g}^1$ and $\boldsymbol{g}^2$, denoted $e_{\boldsymbol{g}^1}(n)$ and $e_{\boldsymbol{g}^2}(n)$ respectively, verify*

$$\max\left(e_{\boldsymbol{g}^1}(n), e_{\boldsymbol{g}^2}(n)\right) \geq \min\left(\frac{1}{128}\exp\left(-\frac{32n}{H_{\text{UNIF}}}\right), \frac{1}{32}\right).$$

Proof Without loss of generality, we assume that n is even. Let $T_k(t)$ be the number of times that arm k has been pulled by the end of round t. Let us consider any fixed learner and a base problem, denoted BASE, with $\nu_1, \ldots, \nu_K$, Bernoulli distributions with means $\mu_1^{\text{BASE}}, \ldots, \mu_K^{\text{BASE}}$ such that $\mu_1^{\text{BASE}} \triangleq 1/2$, and for all $k \in [2, K], \mu_k^{\text{BASE}} \triangleq 1/2 - \Delta_k$ (the gaps specified in the claim of the theorem). The expectation and probability with respect to the learner and the samples of this problem are denoted $\mathbb{E}_{\text{BASE}}$ and $\mathbb{P}_{\text{BASE}}$.

We consider an early period of the game $t = 1, \ldots, n/2$, that we denote as $L \triangleq [1 : n/2]$. The proof could work with any set of $n/2$ rounds that are not necessarily the first of the game but we stick to the early ones to ease the notation. Now we want to identify an arm that the learner will not pull very much in the base problem BASE during the period L. For our learner, during the period L, in any case at least one *fixed* arm among $[2 : K]$ is pulled in expectation less or equal to n/K as

otherwise the total number of pulls is in expectation strictly larger than $(K-1)\times n/K \geq n/2$. We denote this arm as $\overline{k}$. By this construction we have that

$$\mathbb{E}_{\text{BASE}}\left[T_{\overline{k}}\left(\frac{n}{2}\right)\right] \leq \frac{n}{K}. \tag{1}$$

Now we construct the two adversarial problems ADV_1 and ADV_2. These adversarial problems sample $\boldsymbol{g}$ randomly. During phase L, at rounds $t \in L$, for any arm k, $g_{k,t}$ is sampled from the Bernoulli distribution with means $\mu_k^p(t), p \in \{1,2\}$ specifying the problem at hand. The expectation and probability with respect to the learner and the samples of this problem p are denoted by $\mathbb{E}_p$ and $\mathbb{P}_p$.

For problem ADV_1, the distributions of the arms during phase L, $t \in L$, are the same as in the BASE $\mu_k^1(t) \triangleq \mu_k^{\text{BASE}}$ for all $k \in [K]$. After phase L, for $t > n/2$, problem ADV_1 assigns deterministically the following gains, for all $k \neq \overline{k}$, for $t > n/2$, $\boldsymbol{g}_{k,t} = 0$ and $\boldsymbol{g}_{\overline{k},t} = \Delta_{\overline{k}} - \Delta_1$. Therefore, the final expected cumulative gain of arm $\overline{k}$ is

$$\mathbb{E}_1\left[G_{\overline{k},n}\right] = \frac{n}{2}\mu_{\overline{k}}^{\text{BASE}} + \frac{n}{2}\left(\Delta_{\overline{k}} - \Delta_1\right) = \frac{n}{2}\left(\mu_1^{\text{BASE}} - \Delta_1\right)$$

and $\mathbb{E}_1\left[G_{k,n}\right] = n\mu_k^{\text{BASE}}/2$ for all $k \neq \overline{k}$. The problem ADV_2 only differs from the previous one in its stochastic first part (the deterministic second part is the same) and only for arm $\overline{k}$ where, for $t \leq n/2$, where $\mu_{\overline{k}}^2(t) \triangleq \mu_{\overline{k}}^{\text{BASE}} + 2\Delta_1$. Therefore, the final expected cumulative gain of arm $\overline{k}$ is

$$\mathbb{E}_2\left[G_{\overline{k},n}\right] = \frac{n}{2}\left(\mu_1^{\text{BASE}} + \Delta_1\right).$$

Let us consider the events

$$\xi_1 \triangleq \left\{\forall k \in [2:K],\ G_k - \mathbb{E}_1\left[G_k\right] \leq \frac{n\Delta_1}{8}\right\} \cup \left\{\mathbb{E}_1\left[G_1\right] - G_1 \leq \frac{n\Delta_1}{8}\right\} \text{ and}$$
$$\xi_2 \triangleq \left\{\forall k \in [K], k \neq \overline{k},\ G_k - \mathbb{E}_2\left[G_k\right] \leq \frac{n\Delta_1}{8}\right\} \cup \left\{\mathbb{E}_2\left[G_{\overline{k}}\right] - G_{\overline{k}} \leq \frac{n\Delta_1}{8}\right\}.$$

Using a standard Hoeffding argument with a union bound we have that

$$\mathbb{P}_1\left(\xi_1\right) \geq 1 - K\exp\left(-2\left(\frac{\Delta_1}{8}\right)^2\frac{n}{2}\right)$$

and the same result holds for $\mathbb{P}_2\left(\xi_2\right)$. Also note that in the problem p, for any $\boldsymbol{g}$ that is compatible with ξ_p, the gaps associated to $\boldsymbol{g}$ verify $\Delta_k/2 \leq \Delta_k^{\boldsymbol{g}} \leq 3\Delta_k$ for all $k \neq \overline{k}$ and $\Delta_1/2 \leq \Delta_{\overline{k}}^{\boldsymbol{g}} \leq 2\Delta_1$ which gives $4H_{\text{UNIF}} \geq H_{\text{UNIF}}(\boldsymbol{g}) \geq H_{\text{UNIF}}/9$. Also on $\boldsymbol{g} \in \xi_1$, we have $k_{\boldsymbol{g}}^\star = 1$ and on $\boldsymbol{g} \in \xi_2$, we have $k_{\boldsymbol{g}}^\star = \overline{k}$.

Now we prove the following relation between the probability of error in the problem ADV_2, $e_2(n)$, and the probability of successful identification in the problem ADV_1, $1 - e_1(n)$, where in this case, for problem p, $e_p(n)$ is defined here as $e_p(n) \triangleq \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}_p}\left(J_n \neq k_{\boldsymbol{g}}^\star\right)$,

$$e_2(n) + K\exp\left(-\frac{\Delta_1^2 n}{64}\right) \geq \frac{1}{8}\left(\frac{1}{4} - e_1(n)\right)\exp\left(-\frac{32\Delta_1^2 n}{K}\right). \tag{2}$$

We have that

$$\begin{aligned}
e_2(n) + K\exp\left(-\frac{\Delta_1^2 n}{64}\right) &\overset{\textbf{(a)}}{\geq} \mathbb{P}_2\left(J_n \neq \overline{k}\right) \\
&\overset{\textbf{(b)}}{\geq} \mathbb{P}_2\left(J_n = 1\right) \\
&\geq \mathbb{P}_2\left(J_n = 1 \cap T_{\overline{k}}\left(\frac{n}{2}\right) \leq \frac{2n}{K}\right) \\
&\overset{\textbf{(c)}}{\geq} \mathbb{P}_1\left(J_n = 1 \cap T_{\overline{k}}\left(\frac{n}{2}\right) \leq \frac{2n}{K}\right)\frac{1}{8}\exp\left(-\frac{16\Delta_1^2 2n}{K}\right) \\
&\overset{\textbf{(d)}}{\geq} \left(\mathbb{P}_1\left(J_n = 1\right) - \frac{1}{2}\right)\frac{1}{8}\exp\left(-\frac{32\Delta_1^2 n}{K}\right) \\
&\overset{\textbf{(a)}}{\geq} \frac{1}{8}\left(\frac{1}{4} - e_1(n)\right)\exp\left(-\frac{32\Delta_1^2 n}{K}\right),
\end{aligned}$$

where **(a)** is because first, as computed above, for problem p,

$$\mathbb{P}_p\left(k^\star_{\boldsymbol{g}} = 1\right) \geq \mathbb{P}(\xi_p) \geq 1 - K\exp\left(-\frac{\Delta_1^2 n}{64}\right),$$

since; if we denote $1^\star = 1$ (the arm with the highest mean in $p = 1$) and $2^\star = \overline{k}$ (the arm with the highest mean in $p = 2$) and in general $p^\star$, we have that

$$\begin{aligned}
e_p(n) &= 1 - \mathbb{P}_p\left(J_n = k^\star_{\boldsymbol{g}}\right) = 1 - \mathbb{P}_p\left(J_n = k^\star_{\boldsymbol{g}} \cap k^\star_{\boldsymbol{g}} = p^\star\right) - \mathbb{P}_p\left(J_n = k^\star_{\boldsymbol{g}} \cap k^\star_{\boldsymbol{g}} \neq p^\star\right) \\
&\geq 1 - \mathbb{P}_p\left(J_n = p^\star\right) - \mathbb{P}_p\left(k^\star_{\boldsymbol{g}} \neq p^\star\right) \geq 1 - \mathbb{P}_p\left(J_n = p^\star\right) - K\exp\left(-\frac{\Delta_1^2 n}{64}\right).
\end{aligned}$$

Moreover, $K\exp\left(-\Delta_1^2 n/64\right) \leq 1/4$ by an assumption of the theorem. Next, **(b)** is because we have $1 \neq \overline{k}$ by construction and **(c)** uses the change-of-measure argument from Lemma 10. Finally, **(d)** is because

$$\begin{aligned}
\mathbb{P}_1\left(J_n = 1\right) &= \mathbb{P}_1\left(J_n = 1 \cap T_{\overline{k}}\left(\frac{n}{2}\right) \leq \frac{2n}{K}\right) + \mathbb{P}_1\left(J_n = 1 \cap T_{\overline{k}}\left(\frac{n}{2}\right) > \frac{2n}{K}\right) \\
&\leq \mathbb{P}_1\left(J_n = 1 \cap T_{\overline{k}}\left(\frac{n}{2}\right) \leq \frac{2n}{K}\right) + \mathbb{P}_1\left(T_{\overline{k}}\left(\frac{n}{2}\right) > \frac{2n}{K}\right) \\
&\leq \mathbb{P}_1\left(J_n = 1 \cap T_{\overline{k}}\left(\frac{n}{2}\right) \leq \frac{2n}{K}\right) + \frac{1}{2},
\end{aligned}$$

where the last inequality combines Equation 1 and Markov's inequality.

Two cases After proving Equation 2, we finally distinguish two cases.
Case 1) $e_2(n) > \exp\left(-32\Delta_1^2 n/K\right)/64$**:** We claim there exists $\boldsymbol{g} \in \xi_2$, as discussed above, that its best arm is $\overline{k}$ and it possesses a complexity verifying $4H_{\text{UNIF}} \geq H_{\text{UNIF}}(\boldsymbol{g}) \geq H_{\text{UNIF}}/9$, such that the lower bound holds. We prove the statement by contradiction. Indeed if no $\boldsymbol{g} \in \xi_2$ is such that

$e_{\boldsymbol{g}}(n) \geq \exp(-n32\Delta_1^2/K)/128$, then we have

$$\begin{aligned}
e_2(n) &= \mathbb{P}_{\boldsymbol{g}\sim \mathrm{ADV}_2}\left(J_n \neq k^\star_{\boldsymbol{g}}\right) \\
&\leq \mathbb{P}\left(J_n \neq k^\star_{\boldsymbol{g}} \cap \boldsymbol{g} \in \xi_2\right) + \mathbb{P}\left(J_n \neq k^\star_{\boldsymbol{g}} \cap \boldsymbol{g} \notin \xi_2\right) \\
&\leq \mathbb{P}\left(J_n \neq k^\star_{\boldsymbol{g}} \cap \boldsymbol{g} \in \xi_2\right) + \mathbb{P}\left(\boldsymbol{g} \notin \xi_2\right) \\
&\leq \frac{1}{128}\exp\left(-\frac{n32\Delta_1^2}{K}\right) + K\exp\left(-\frac{\Delta_1^2 n}{64}\right) \\
&= \frac{1}{128}\exp\left(-\frac{n32\Delta_1^2}{K}\right) + K\exp\left(-\frac{\Delta_1^2 n}{128}\right)\exp\left(-\frac{\Delta_1^2 n}{128}\right) \\
&\overset{\textbf{(a)}}{\leq} \frac{1}{128}\exp\left(-\frac{n32\Delta_1^2}{K}\right) + \frac{1}{128}\exp\left(-\frac{n\Delta_1^2}{128}\right) \\
&\overset{\textbf{(b)}}{\leq} \frac{1}{128}\exp\left(-\frac{n32\Delta_1^2}{K}\right) + \frac{1}{128}\exp\left(-\frac{n32\Delta_1^2}{K}\right),
\end{aligned}$$

where **(a)** is because $K\exp\left(-n\Delta_1^2/128\right) \leq 1/128$ and **(b)** because $K > 32\times 128$ by assumptions. This is a contradiction with the original assumption of that case.

Case 2) $e_2(n) \leq \exp(-32\Delta_1^2 n/K)/64$**:** We first want to prove that this assumption gives

$$e_1(n) \geq \frac{1}{16}, \tag{3}$$

that we prove by first using Equation 2 which gives

$$\frac{1}{64}\exp\left(-\frac{32\Delta_1^2 n}{K}\right) + K\exp\left(-\frac{\Delta_1^2 n}{64}\right) \geq \frac{1}{8}\left(\frac{1}{4} - e_1(n)\right)\exp\left(-\frac{32\Delta_1^2 n}{K}\right),$$

and hence

$$\frac{1}{8} + 8K\frac{\exp\left(-\frac{\Delta_1^2 n}{64}\right)}{\exp\left(-\frac{32\Delta_1^2 n}{K}\right)} \geq \frac{1}{4} - e_1(n).$$

Therefore, we have

$$\begin{aligned}
e_1(n) &\geq \frac{1}{8} - 8K\exp\left(\frac{32\Delta_1^2 n}{K} - \frac{\Delta_1^2 n}{64}\right) \\
&\overset{\textbf{(a)}}{\geq} \frac{1}{8} - 8K\exp\left(-\frac{\Delta_1^2 n}{128}\right) \\
&\overset{\textbf{(b)}}{\geq} \frac{1}{16},
\end{aligned}$$

where **(a)** is because $K > 32\times 128$ and **(b)** is because $K\exp\left(-n\Delta_1^2/128\right) \leq 1/128$ by assumptions. We now claim that there exists at least one $\boldsymbol{g} \in \xi_1$ with best arm 1 such that $e_{\boldsymbol{g}}(n) \geq 1/32$.

The proof is by contradiction. Let us assume that for all $\boldsymbol{g} \in \xi_1$, $e_{\boldsymbol{g}}(n) < 1/32$. Then, we have

$$\begin{aligned}
e_1(n) &= \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}_1}\left(J_n \neq k^\star_{\boldsymbol{g}}\right) \\
&= \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}_1}\left(J_n \neq k^\star_{\boldsymbol{g}} | \, \boldsymbol{g} \in \xi_1\right) P\left(\xi_1\right) + \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}_1}\left(J_n \neq k^\star_{\boldsymbol{g}} | \, \boldsymbol{g} \notin \xi_1\right) P\left(\neg\xi_1\right) \\
&\leq \frac{1}{32} \times 1 + 1 \times K \exp\left(-\frac{\Delta_1^2 n}{64}\right) \\
&\leq \frac{1}{32} \times 1 + 1 \times \frac{1}{128} \\
&< \frac{1}{16},
\end{aligned}$$

which contradicts Equation 3. ■

Appendix E. Lower bound in the best of both worlds

Theorem 4 (Lower bound for the BOB challenge) *For any class problem $\boldsymbol{\Delta}_4$, for any learner, there exists an i.i.d. stochastic problem* STO *with complexity H_{BOB} and there exists an adversarial problem $\boldsymbol{g}$ such that for any n satisfying $K \exp\left(-\Delta_1^2 n / 32\right) \leq 1/32$, if the probability of error of the learner on* STO *satisfies*

$$e_{\text{STO}}(n) \leq \frac{1}{64} \exp\left(-\frac{2048 n}{H_{\text{BOB}}}\right),$$

then, in the adversarial problem, the learner suffers a constant error,

$$e_{\text{ADV}(\boldsymbol{g})}(n) \geq \frac{1}{16}.$$

Proof Let $T_k(t)$ be the number of times that arm k has been pulled by the end of round t. Let us consider any fixed learner. Let us consider a base problem, denoted BASE, with $\nu_1, \ldots, \nu_K$, Bernoulli distributions with mean $\mu_1^{\text{BASE}}, \ldots, \mu_K^{\text{BASE}}$ such that $\mu_1^{\text{BASE}} \triangleq 1/2$, and for all $k \in [2, K]$, $\mu_k^{\text{BASE}} \triangleq 1/2 - \Delta_k$ (the gaps specified in the claim of the theorem). The expectation and probability with respect to the learner and the samples of this problem are $\mathbb{E}_{\text{BASE}}$ and $\mathbb{P}_{\text{BASE}}$.

Here, $i \in [2 : K]$ denotes the rank of a suboptimal arm in the base problem. Next, we consider a constant $a_i \leq 1$. We also consider an early period of the game $t = 1, \ldots, n_i \triangleq \lceil n a_i \rceil$, that we denote L_i. The proof could work with any set of n_i rounds that are not necessarily the first of the game but we stick to the early ones to ease the notation. Now we want to identify an arm with a small gap that the learner will not pull very much in the base problem BASE during the period L_i. From our learner, during the period L_i, in any case at most $i - 2$ arms among $[2 : K]$ will, in expectation, be pulled strictly more than $2n_i/i$ as otherwise the total number of pulls is strictly larger than $(i-1) \times 2n_i/i \geq n_i$. Therefore, we have at least $K - 1 - (i-2) = K - i + 1$ arms included in the set $[2 : K]$, and that form a set noted S, that are pulled in expectation less than $2n_i/i$. Among these arms, let us consider arm $\overline{k} \triangleq \text{argmax}_{k \in S} \mu_k$ with highest mean. By construction we have

$$\mathbb{E}_{\text{BASE}}\left[T_{\overline{k}}(n_i)\right] \leq \frac{2n_i}{i}. \tag{4}$$

Note that by construction, we have also that $\Delta_{\overline{k}} \leq \Delta_i$, because otherwise it would mean that $\mu_{\overline{k}} < \mu_i$ and so there would exist at most $K-i-1$ arms with lower means than $\overline{k}$. This contradicts the fact that $\overline{k}$ has the highest mean among $K-i+1$ arms.

Now we construct an i.i.d. stochastic problem, denoted STO, where the distribution of the arms are the same as in the BASE problem except for arm $\overline{k}$, $\mu_k^{\text{STO}} \triangleq \mu_k^{\text{BASE}}$ for all $k \neq \overline{k}$. We set in STO, $\mu_{\overline{k}}^{\text{STO}} \triangleq 1/2 + \Delta_1 \frac{\Delta_{\overline{k}}}{\Delta_i}/2$.

This means that the best arm in the STO is the arm $\overline{k}$, $k^\star_{\text{STO}} \triangleq \overline{k}$. Also note that the gaps in the STO problem verify $\Delta_k \leq \Delta_k^{\text{STO}} = \Delta_k + \Delta_1 \frac{\Delta_{\overline{k}}}{\Delta_i}/2 \leq 2\Delta_k$ for all $k \in [2, \overline{k}-1] \cup [\overline{k}+1 : K]$.

Also, $\Delta_1^{\text{STO}} = \Delta_{\overline{k}}^{\text{STO}} = \Delta_1 \frac{\Delta_{\overline{k}}}{\Delta_i}/2$. Therefore,

$$H_{\text{BOB}}^{\text{STO}} = \frac{1}{\Delta_{\overline{k}}^{\text{STO}}} \max_k \frac{k}{\Delta_k^{\text{STO}}} = \frac{2\Delta_i}{\Delta_1 \Delta_{\overline{k}}} \max_k \frac{k}{\Delta_k^{\text{STO}}} \overset{\textbf{(a)}}{\leq} \frac{2\Delta_i}{\Delta_1 \Delta_{\overline{k}}} \max_k \frac{k}{\Delta_k} \frac{2\Delta_i}{\Delta_{\overline{k}}}, \tag{5}$$

where **(a)** is because, between the BASE and the STO, as we have seen, the gaps verify for all k $\Delta_k^{\text{STO}} \geq \Delta_k \Delta_{\overline{k}}/(2\Delta_i)$.

The expectation and probability with respect to the learner and the samples of this problem are denoted $\mathbb{E}_{\text{STO}}$ and $\mathbb{P}_{\text{STO}}$.

The second bandit problem is the adversarial one, denoted ADV. This adversarial problem samples $\boldsymbol{g}$ randomly. At round $t \in [n]$, for arm k, $g_{k,t}$ is sampled from the Bernoulli distribution with mean $\mu_k^{\text{ADV}}(t)$. For all t and $k \neq \overline{k}$, ADV follows the BASE problem: $\mu_k^{\text{ADV}}(t) \triangleq \mu_k^{\text{BASE}}$. For arm $\overline{k}$, until the end of phase L_i, for all t with $1 \leq t \leq n_i$, $\mu_{\overline{k}}^{\text{ADV}}(t) \triangleq \mu_{\overline{k}}^{\text{BASE}}$ and then a switch happens, for $n_i < t \leq n$, arm $\overline{k}$ possesses the same distributions as in the STO problem, $\mu_{\overline{k}}^{\text{ADV}}(t) \triangleq \mu_{\overline{k}}^{\text{STO}}$. The expectation and probability with respect to the learner and the samples of this problem are denoted $\mathbb{E}_{\text{ADV}}$ and $\mathbb{P}_{\text{ADV}}$.

Let us now study the identity of the best arm in ADV. We want to show that, with high probability, the best arm in the ADV is arm 1 if we have $a_i \geq \Delta_1/\Delta_i$. We denote the expected cumulative gain in ADV of each arm $k \in [K]$ as $M_k \triangleq \sum_{t=1}^n \mu_k^{\text{ADV}}(t)$. For arm $\overline{k}$, we have

$$\begin{aligned}
M_{\overline{k}} &= \sum_{t=1}^{n_i} \mu_{\overline{k}}^{\text{ADV}}(t) + \sum_{t=n_i+1}^{n} \mu_{\overline{k}}^{\text{ADV}}(t) \\
&= n_i \mu_{\overline{k}}^{\text{BASE}} + (n-n_i)\mu_{\overline{k}}^{\text{STO}} \\
&= n_i \left(\frac{1}{2} - \Delta_{\overline{k}}\right) + (n-n_i)\left(\frac{1}{2} + \frac{\Delta_1}{2}\frac{\Delta_{\overline{k}}}{\Delta_i}\right) \\
&= \frac{n}{2} - n_i\Delta_{\overline{k}} + n\frac{\Delta_1}{2}\frac{\Delta_{\overline{k}}}{\Delta_i} - n_i\frac{\Delta_1}{2}\frac{\Delta_{\overline{k}}}{\Delta_i} \\
&\leq \frac{n}{2} - n_i\Delta_{\overline{k}} + n\frac{\Delta_1}{2}\frac{\Delta_{\overline{k}}}{\Delta_i} \\
&\leq \frac{n}{2} - a_i n\Delta_{\overline{k}} + \frac{n\Delta_1}{2}\frac{\Delta_{\overline{k}}}{\Delta_i} \\
&\leq \frac{n}{2} - n\Delta_1\frac{\Delta_{\overline{k}}}{\Delta_i} + \frac{n\Delta_1}{2}\frac{\Delta_{\overline{k}}}{\Delta_i} \\
&= \frac{n}{2} - \frac{n\Delta_1}{2}\frac{\Delta_{\overline{k}}}{\Delta_i}.
\end{aligned}$$

For all $k \in [K]$, $k \neq \overline{k}$ and $k \neq 1$, we have $M_k = n/2 - n\Delta_k$. Furthermore, $M_1 = n/2$. Let us consider the event $\xi = \left\{\forall k \in [K], |G_k - M_k| \leq n\Delta_1/8\frac{\Delta_{\overline{k}}}{\Delta_i}\right\}$. Using a standard Hoeffding argument with a union bound we have that $\mathbb{P}(\xi) \geq 1 - K\exp\left(-\Delta_1^3 n/32\right)$.

here i put a power of 3 to be safe and i ttink it doesnt matter in the proof because we put it as an hypothesis of the theorem that n is large enough

Then, in the ADV problem, for any $\boldsymbol{g}$ that is compatible with ξ, the gaps associated with $\boldsymbol{g}$ verify $\Delta_k/2 \leq \Delta_k^{\boldsymbol{g}} \leq 2\Delta_k$ for all $k \neq \overline{k}$ and $k \neq 1$. And $\Delta_1\frac{\Delta_{\overline{k}}}{\Delta_i}/4 \leq \Delta_{\overline{k}}^{\boldsymbol{g}} \leq \Delta_1\frac{\Delta_{\overline{k}}}{\Delta_i}$. And $\Delta_1\frac{\Delta_{\overline{k}}}{\Delta_i}/4 \leq \Delta_{\overline{k}}^{\boldsymbol{g}} \leq \Delta_1\frac{\Delta_{\overline{k}}}{\Delta_i}$

Note that therefore the only difference between the two problems STO and ADV is for arm $\overline{k}$ during phase L_i.

If we have $a_i \geq \Delta_1/\Delta_i$, then with high probability, the respective best arms in problem STO and problem ADV are different, i.e., $k^\star_{\text{ADV}} \neq k^\star_{\text{STO}}$. That is what we assume for the rest of the proof. Indeed, we want to use the fact that the two models are hard to differentiate from the learner point of view with a certain probability and that then the learner has to either choose to recommend $k^\star_{\text{ADV}}$ or $k^\star_{\text{STO}}$, which are different, and therefore possibly suffer a mistake.

Then we prove the following relation between the probability of error in the stochastic problem $e_{\text{STO}}(n)$ to the probability of successful identification in the adversarial problem $1 - e_{\text{ADV}}(n)$: where $e_{\text{ADV}}(n)$ is defined here as $e_{\text{ADV}}(n) \triangleq \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(J_n \neq k^\star_{\boldsymbol{g}}\right)$,

$$e_{\text{STO}}(n) \geq \frac{1}{8}\left(\frac{1}{4} - e_{\text{ADV}}(n)\right)\exp\left(-\frac{64\Delta_{\overline{k}}^2 n_i}{i}\right). \tag{6}$$

To obtain Equation 6, we write

$$\begin{aligned}
e_{\text{STO}}(n) &= \mathbb{P}_{\text{STO}}\left(J_n \neq \overline{k}\right) \\
&\overset{\text{(a)}}{\geq} \mathbb{P}_{\text{STO}}\left(J_n = 1\right) \\
&\geq \mathbb{P}_{\text{STO}}\left(J_n = 1 \cap T_{\overline{k}}(n_i) \leq \frac{4n_i}{i}\right) \\
&\overset{\text{(b)}}{\geq} \mathbb{P}_{\text{ADV}}\left(J_n = 1 \cap T_{\overline{k}}(n_i) \leq \frac{4n_i}{i}\right)\frac{1}{8}\exp\left(-\frac{16\Delta_{\overline{k}}^2 4n_i}{i}\right) \\
&\overset{\text{(c)}}{\geq} \left(\mathbb{P}_{\text{ADV}}\left(J_n = 1\right) - \frac{1}{2}\right)\frac{1}{8}\exp\left(-\frac{64\Delta_{\overline{k}}^2 n_i}{i}\right) \\
&\overset{\text{(d)}}{\geq} \frac{1}{8}\left(\frac{1}{4} - e_{\text{ADV}}(n)\right)\exp\left(-\frac{64\Delta_{\overline{k}}^2 n_i}{i}\right),
\end{aligned}$$

where **(a)** is because we have $1 \neq \overline{k}$ by construction **(b)** uses the change-of-measure argument from Lemma 10. Step **(c)** is because

$$\begin{aligned}
\mathbb{P}_{\text{ADV}}\left(J_n = 1\right) &= \mathbb{P}_{\text{ADV}}\left(J_n = 1 \cap T_{\overline{k}}(n_i) \leq \frac{4n_i}{i}\right) + \mathbb{P}_{\text{ADV}}\left(J_n = 1 \cap T_{\overline{k}}(n_i) > \frac{4n_i}{i}\right) \\
&\leq \mathbb{P}_{\text{ADV}}\left(J_n = 1 \cap T_{\overline{k}}(n_i) \leq \frac{4n_i}{i}\right) + \mathbb{P}_{\text{ADV}}\left(T_{\overline{k}}(n_i) > \frac{4n_i}{i}\right) \\
&\leq \mathbb{P}_{\text{ADV}}\left(J_n = 1 \cap T_{\overline{k}}(n_i) \leq \frac{4n_i}{i}\right) + \frac{1}{2},
\end{aligned}$$

where the last inequality combines Equation 4 and a Markov inequality. Step **(d)** is because first, as computed above, $\mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(k^\star_{\boldsymbol{g}} = 1\right) \geq \mathbb{P}(\xi) \geq 1 - K\exp\left(-\Delta_1^3 n/32\right)$ and therefore, we have

$$\begin{aligned}
e_{\text{ADV}}(n) &= 1 - \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(J_n = k^\star_{\boldsymbol{g}}\right) \\
&= 1 - \mathbb{P}_{\text{ADV}}\left(J_n = k^\star_{\boldsymbol{g}} \cap k^\star_{\boldsymbol{g}} = 1\right) - \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(J_n = k^\star_{\boldsymbol{g}} \cap k^\star_{\boldsymbol{g}} \neq 1\right) \\
&\geq 1 - \mathbb{P}_{\text{ADV}}\left(J_n = 1\right) - \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(k^\star_{\boldsymbol{g}} \neq 1\right) \\
&\geq 1 - \mathbb{P}_{\text{ADV}}\left(J_n = 1\right) - K\exp\left(-\frac{\Delta_1^3 n}{32}\right) \\
&\geq 1 - \mathbb{P}_{\text{ADV}}\left(J_n = 1\right) - \frac{1}{4},
\end{aligned}$$

where $K\exp\left(-\Delta_1^3 n/32\right) \leq 1/4$ holds by assumption of the theorem. Having just proved Equation 6, we proceed with the rest of the proof. In order to maximize the lower bound we maximize n_i by setting $a_i \triangleq \Delta_1/\Delta_i$. Then again to maximize the lower bound, we finally choose $i \triangleq \text{argmax}_{k\in[K]}(k/\Delta_k)$. Rewriting Equation 5 given the choice $i \triangleq \text{argmax}_{k\in[K]}(k/\Delta_k)$, we have

$$H^{\text{STO}}_{\text{BOB}} \leq \frac{2\Delta_i}{\Delta_1\Delta_{\overline{k}}}\frac{i}{\Delta_i}\frac{2\Delta_i}{\Delta_{\overline{k}}} = \frac{2i}{\Delta_1}\frac{2\Delta_i}{\Delta^2_{\overline{k}}}, \tag{7}$$

$$\begin{aligned}
\frac{1}{64}\exp\left(-\frac{2048n}{H^{\text{STO}}_{\text{BOB}}}\right) &\leq \frac{1}{64}\exp\left(-\frac{256n\Delta_1\Delta^2_{\overline{k}}}{i\Delta_i}\right) \\
&= \frac{1}{64}\exp\left(-\frac{128\Delta^2_{\overline{k}}}{i}\frac{2\Delta_1 n}{\Delta_i}\right) \\
&\leq \frac{1}{64}\exp\left(-\frac{128\Delta^2_{\overline{k}}}{i}\left(\left\lceil\frac{2\Delta_1 n}{\Delta_i}\right\rceil - 1\right)\right) \\
&\leq \frac{1}{64}\exp\left(-\frac{64\Delta^2_{\overline{k}} n_i}{i}\right).
\end{aligned}$$

Therefore, if $e_{\text{STO}}(n) \leq 1/64\exp\left(-2048n/H^{\text{STO}}_{\text{BOB}}\right)$ then using the inequality above, we get that $e_{\text{STO}}(n) \leq \exp\left(-64\Delta^2_{\overline{k}} n_i/i\right)/64$. Finally, using the inequality in Equation 6, we have that if $e_{\text{STO}}(n) \leq \exp\left(-64\Delta^2_{\overline{k}} n_i/i\right)/64$ then $e_{\text{ADV}}(n) \geq 1/8$.

We now claim that there exists at least one $\boldsymbol{g} \in \xi$ such that $e_{\boldsymbol{g}}(n) \geq 1/16$. The proof is by contradiction. Let us assume that for all $\boldsymbol{g} \in \xi$, $e_{\boldsymbol{g}}(n) < 1/16$. Then, we have

$$\begin{aligned}
e_{\text{ADV}}(n) &= \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(J_n \neq k^\star_{\boldsymbol{g}}\right) \\
&= \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(J_n \neq k^\star_{\boldsymbol{g}} | \, \boldsymbol{g} \in \xi\right) P(\xi) + \mathbb{P}_{\boldsymbol{g}\sim\text{ADV}}\left(J_n \neq k^\star_{\boldsymbol{g}} | \, \boldsymbol{g} \notin \xi\right) P(\neg\xi) \\
&\leq \frac{1}{16} \times 1 + 1 \times K \exp\left(-\frac{\Delta_1^2 n}{32}\right) \\
&\leq \frac{1}{16} \times 1 + 1 \times \frac{1}{32} \\
&< \frac{1}{8},
\end{aligned}$$

which is a contradiction. ■

Appendix F. Upper bound in best of both worlds

Lemma 11 *Let ξ_p be the events defined in Equation 9 for all $p \in [2:K]$. On the conjunction of events $\cap_{p=i+1}^{K+1}\xi_p$ and for $i \in [2:K]$, in an i.i.d. stochastic environment, and complexity $H_{\texttt{P1}}$, playing* P1*, given two distinct arms $j \in [i:K]$, $k \in [K]$ and a round $t \geq n_i$ such that $\mu_1 - \mu_k < \Delta_i/2$, we have*

$$\mathbb{P}\left(\widetilde{G}_{k,t} \leq \widetilde{G}_{j,t}\right) \leq 2\exp\left(-\frac{n}{128 H_{\texttt{P1}}}\right).$$

Proof We prove that for any proportions of rounds $\boldsymbol{a}$, we have

$$\mathbb{P}\left(\widetilde{G}_{k,t} \leq \widetilde{G}_{j,t}\right) \leq 2\exp\left(-\frac{n}{128 H_{\texttt{P1}}(\boldsymbol{a})}\right).$$

Then, the claim of the Lemma 11 comes from $H_{\texttt{P1}} = \min_{\boldsymbol{a}\in\boldsymbol{A}} H_{\texttt{P1}}(\boldsymbol{a})$. First, notice that

$$\mu_k - \mu_j = \mu_k - \mu_1 + \mu_1 - \mu_j \overset{\textbf{(a)}}{>} -\frac{\Delta_j}{2} + \mu_1 - \mu_j = \frac{\Delta_j}{2} > 0,$$

where **(a)** is because by the assumption of the lemma, $\mu_1 - \mu_k < \Delta_i/2 \leq \Delta_j/2$. We decompose

$$\begin{aligned}
\mathbb{P}\left(\widetilde{G}_{k,t} \leq \widetilde{G}_{j,t}\right) &\overset{\textbf{(c)}}{\leq} \mathbb{P}\left(t\mu_k - \widetilde{G}_{k,t} \geq t\frac{\mu_k - \mu_j}{2}\right) + \mathbb{P}\left(\widetilde{G}_{j,t} - t\mu_j \geq t\frac{\mu_k - \mu_j}{2}\right) \\
&\overset{\textbf{(d)}}{\leq} \mathbb{P}\left(t\mu_k - \widetilde{G}_{k,t} > \frac{t\Delta_j}{4}\right) + \mathbb{P}\left(\widetilde{G}_{j,t} - t\mu_j > \frac{t\Delta_j}{4}\right), \qquad (8)
\end{aligned}$$

where **(c)** is because $\mu_k - \mu_j > 0$ and **(d)** is because $\mu_k - \mu_j > \Delta_j/2$.

We now bound the two terms in Equation 8. To bound the *second term* in Equation 8 we have for all $t' \in [n]$, $|\widetilde{\boldsymbol{g}}_{j,t'} - \boldsymbol{g}_{j,t'}| \leq K\overline{\log}K$ as $\boldsymbol{p}_{j,t'} \geq 1/(K\overline{\log}K)$. We define arm j^+ so that $j^+ + 1$ is the arm with the largest mean among the ones that have the at least twice the gap of the gap of j, $j^+ + 1 \triangleq \underset{p':\mu_1-\mu_j<\Delta_{p'}/2}{\arg\max} \mu_{p'}$. Note that as $j \geq i > 1$, we have $j^+ + 1 > j$ and therefore

$j^+ \geq j \geq i$. We now bound the cumulative variance $\sum_{t'=1}^{t} \sigma^2_{\widetilde{\boldsymbol{g}}_{j,t'} - \boldsymbol{g}_{j,t'}}$ for the mean estimator of arm j at round t',

$$\begin{aligned}
\sum_{t'=1}^{t} \sigma^2_{\widetilde{\boldsymbol{g}}_{j,t'} - \boldsymbol{g}_{j,t'}} &\overset{\textbf{(e')}}{=} \sum_{\ell=j^+}^{K} \sum_{t'=n_{\ell+1}+1}^{n_\ell} \sigma^2_{\widetilde{\boldsymbol{g}}_{j,t'} - \boldsymbol{g}_{j,t'}} + \sum_{t'=n_{j^+}+1}^{t} \sigma^2_{\widetilde{\boldsymbol{g}}_{j,t'} - \boldsymbol{g}_{j,t'}} \\
&\overset{\textbf{(e)}}{\leq} \sum_{\ell=j^+}^{K} \sum_{t'=n_{\ell+1}+1}^{n_\ell} \ell \overline{\log} K + \sum_{t'=n_{j^+}+1}^{t} j^+ \overline{\log} K \\
&= \sum_{\ell=j^+}^{K} (n_\ell - n_{\ell+1}) \ell \overline{\log} K + (t - n_{j^+}) j^+ \overline{\log} K,
\end{aligned}$$

where **(e')** is because $t \geq n_i \geq n_{j^+}$ and **(e)** is because we are on the conjunction of events $\cap_{p=i+1}^{K+1} \xi_p$. Therefore, as for each round t' in the round interval $[n_{\ell+1}+1 : n_\ell]$ with $\ell \in [j^+ : K]$, we verify $\mu_1 - \mu_j < \Delta_{j^+ +1}/2 \leq \Delta_{\ell+1}/2$, then we have $\widetilde{\langle j \rangle}_{t'} \leq \ell$. For $t' \in [n_{j^+}+1 : t]$, we use the fact that event $\xi_{j^+ +1}$ holds for arm j by construction of the events $\{\xi_i\}_i$ and therefore $\forall t' > n_{j^+} \geq n_{j^+ +1}$, $\widetilde{\langle j \rangle}_{t'} < j^+ \leq j^+ + 1$ and we can bound the variance.

By applying the Bernstein inequality for martingale differences we have

$$\begin{aligned}
\mathbb{P}\left(\widetilde{G}_{j,t} - t\mu_j > \frac{t\Delta_j}{4} \right) &\overset{\textbf{(f)}}{\leq} \mathbb{P}\left(\widetilde{G}_{j,t} - t\mu_j > \frac{t\Delta_{j^+}}{8} \right) \\
&\leq \exp\left(- \frac{\left(t\Delta_{j^+}/8\right)^2}{2\sum_{t'=1}^{t} \sigma^2_{\widetilde{\boldsymbol{g}}_{j,t'} - \boldsymbol{g}_{j,t'}} + \frac{2}{3} K \overline{\log}(K) \frac{t\Delta_{j^+}}{8}} \right) \\
&\leq \exp\left(- \frac{\left(t\Delta_{j^+}\right)^2/64}{2\sum_{\ell=j^+}^{K} (n_\ell - n_{\ell+1}) \ell \overline{\log} K + (t - n_{j^+}) j^+ \overline{\log} K + \frac{1}{12} K \overline{\log}(K) t \Delta_{j^+}} \right) \\
&\overset{\textbf{(g)}}{\leq} \exp\left(- \frac{\left(n_{j^+}\Delta_{j^+}\right)^2/64}{2\sum_{\ell=j^+}^{K} (n_\ell - n_{\ell+1}) \ell \overline{\log} K + \frac{1}{12} K \overline{\log}(K) n_{j^+} \Delta_{j^+}} \right),
\end{aligned}$$

where **(f)** is because $\Delta_{j^+}/2 \leq \Delta_j$ by construction and **(g)** is because the exponential term is a decreasing function of t and $t \geq n_i \geq n_{j^+}$.

To bound the *first term* of Equation 8 we use similar arguments. We have for all $t' \in [n]$, $|\widetilde{\boldsymbol{g}}_{k,t'} - \boldsymbol{g}_{k,t'}| \leq K\overline{\log} K$ as $\boldsymbol{p}_{j,t'} \geq 1/(K\overline{\log} K)$. We get

$$\begin{aligned}
\sum_{t'=1}^{t} \sigma^2_{\widetilde{\boldsymbol{g}}_{k,t'} - \boldsymbol{g}_{k,t'}} &= \sum_{\ell=i}^{K} \sum_{t'=n_{\ell+1}+1}^{n_\ell} \sigma^2_{\widetilde{\boldsymbol{g}}_{k,t'} - \boldsymbol{g}_{k,t'}} + \sum_{t'=n_i+1}^{t} \sigma^2_{\widetilde{\boldsymbol{g}}_{k,t'} - \boldsymbol{g}_{k,t'}} \\
&\overset{\textbf{(e)}}{\leq} \sum_{\ell=i}^{K} \sum_{t'=n_{\ell+1}+1}^{n_\ell} \ell \overline{\log} K + \sum_{t'=n_i+1}^{t} i \overline{\log} K \\
&= \sum_{\ell=i}^{K} (n_\ell - n_{\ell+1}) \ell \overline{\log} K + (t - n_i) i \overline{\log} K,
\end{aligned}$$

where **(e)** is because we are on the conjunction of events $\cap_{p=i+1}^{K}\xi_p$. Therefore, as for each round t' in the round interval $[n_{\ell+1}+1 : n_\ell]$ with $\ell \in [i : K]$, we verify $\mu_1 - \mu_k < \Delta_{i+1}/2 \leq \Delta_{\ell+1}/2$, then we have $\widetilde{\langle k \rangle}_{t'} < \ell + 1$. For $t' \in [n_i + 1 : t]$, we use the fact that event ξ_{i+1} holds for arm k by construction and therefore $\forall t' \geq n_{i+1} \geq n_i$, we have $\widetilde{\langle k \rangle}_{t'} \leq i$ and we can bound the variance. We apply the Bernstein inequality for martingale differences again to get

$$\begin{aligned}
\mathbb{P}\left(\widetilde{G}_{k,t} - t\mu_k > \frac{t\Delta_j}{4}\right) &\overset{\textbf{(f)}}{\leq} \mathbb{P}\left(\widetilde{G}_{k,t} - t\mu_k > \frac{t\Delta_i}{4}\right) \\
&\leq \exp\left(\frac{-(t\Delta_i/4)^2}{2\sum_{t'=1}^{t}\sigma^2_{\widetilde{\boldsymbol{g}}_{j,t'}-\boldsymbol{g}_{j,t'}} + \frac{2}{3}K\overline{\log}(K)\frac{t\Delta_i}{4}}\right) \\
&\leq \exp\left(\frac{-(t\Delta_i)^2/16}{2\sum_{\ell=i}^{K}(n_\ell - n_{\ell+1})\ell\overline{\log}K + (t-n_i)i\overline{\log}K + \frac{1}{6}K\overline{\log}(K)t\Delta_i}\right) \\
&\overset{\textbf{(g)}}{\leq} \exp\left(\frac{-(n_i\Delta_i)^2/16}{2\sum_{\ell=i}^{K}(n_\ell - n_{\ell+1})\ell\overline{\log}K + \frac{1}{6}K\overline{\log}(K)n_i\Delta_i}\right),
\end{aligned}$$

where **(f)** is because $\Delta_i \leq \Delta_j$ by construction and **(g)** is because the exponential term is a decreasing function of t and $t \geq n_i \geq n_{j^+}$. ■

Lemma 12 *Let ξ_p be the events defined in Equation 9 for all $p \in [2 : K]$. In an i.i.d. stochastic environment and complexity $H_{\texttt{P1}}$ playing* P1*, we have for all $i \in [2 : K]$,*

$$\mathbb{P}\left(\xi_i^c \,\middle|\, \bigcap_{p=i+1}^{K+1}\xi_p\right) \leq 2K^2 n \exp\left(-\frac{n}{128H_{\texttt{P1}}}\right).$$

Proof Let us consider one arm $k \in [K]$ and a round $t > n_i$ such that $\mu_1 - \mu_k < \Delta_i/2$. We bound the probability that $\widetilde{\langle k \rangle}_t \geq i$ as

$$\begin{aligned}
\mathbb{P}\left(\widetilde{\langle k \rangle}_t \geq i\right) &\overset{\textbf{(a)}}{\leq} \mathbb{P}\left(\exists j \in [i : K],\ \widetilde{G}_{k,t-1} \leq \widetilde{G}_{j,t-1}\right) \\
&\leq \sum_{j=i}^{K}\mathbb{P}\left(\widetilde{G}_{k,t-1} \leq \widetilde{G}_{j,t-1}\right) \\
&\overset{\textbf{(b)}}{\leq} \sum_{j=i}^{K} 2\exp\left(-\frac{n}{128H_{\texttt{P1}}}\right),
\end{aligned}$$

where **(a)** is because if $\forall j \in [i : K],\ \widetilde{G}_{k,t} > \widetilde{G}_{j,t}$, then we have $\widetilde{\langle k \rangle}_t < i$, **(b)** is using Lemma 11 with $t' = t - 1$ and we have $t' = t - 1 > n_i - 1 \geq n_i$. Using union bounds on the arms in $k \in [K]$ and the rounds t, we get the claim of the lemma. ■

Theorem 6 (Upper bounds for P1) *For any stochastic problem* STO *with complexity* $H_{\texttt{P1}}$ *and for any* $\boldsymbol{g}$ *fixed by an oblivious adversary with complexity* $H_{\text{UNIF}}(\boldsymbol{g})$*, the probabilities of error of* P1*, denoted* $e_{\text{STO}}(n)$ *and* $e_{\text{ADV}(\boldsymbol{g})}(n)$ *in their respective environment, for any* n*, simultaneously verify*

$$e_{\text{STO}}(n) \leq 2K^3 n \exp\left(-\frac{n}{128 H_{\texttt{P1}}}\right) \quad \textit{and} \quad e_{\text{ADV}(\boldsymbol{g})}(n) \leq K \exp\left(-\frac{3n}{40\overline{\log}(K) H_{\text{UNIF}}(\boldsymbol{g})}\right).$$

Proof We consider two cases separately.

The i.i.d. stochastic case We place ourselves in the i.i.d. stochastic setting described in Section 4. To ease the notation and without loss of generality, we assume that the arms are sorted by their means so that arm 1 is the best, $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$, and $\Delta_1 = \Delta_2 \leq \ldots \leq \Delta_K$.

Let us consider any rounds verifying $n_2 = n \geq n_3 \geq \ldots \geq n_{K+1} = 0$. Intuitively, n_i is a round after which, for $t \geq n_i$, we expect P1 to have well ranked any arm k with a gap smaller than half the gap of arm i, in the sense that $\widetilde{\langle k \rangle}_t < i$, if $\mu_1 - \mu_k \leq \Delta_i/2$. For $i \in [2 : K+1]$, we define the following event ξ_i,

$$\xi_i \triangleq \left\{ \forall t > n_i, \forall k \in [K] : \mu_1 - \mu_k < \frac{\Delta_i}{2} \implies \widetilde{\langle k \rangle}_t < i \right\}. \tag{9}$$

Note that as the ranks $\widetilde{\langle k \rangle}_t$ are integers, $\widetilde{\langle k \rangle}_t < i$ is equivalent to $\widetilde{\langle k \rangle}_t \leq i - 1$. We initialize the sequence by defining $\Delta_{K+1} = 3\Delta_K$ and then we have that event ξ_{K+1} is always true.

If ξ_2 holds, the algorithm P1 makes no mistake as $\widetilde{\langle k \rangle}_{n+1} = 1$ and the returned arm is $J_n = 1$. More generally, we say for $i \in [2 : K]$, that if ξ_i does not hold, or equivalently if its complement ξ_i^c holds, then the algorithm makes a mistake at stage i. We now bound the probability of an event A with respect to the probability of mistake at a stage j, $\mathbb{P}(\xi_j^c)$ as

$$\mathbb{P}(A) = \mathbb{P}\left(A \cap \xi_j^c\right) + \mathbb{P}\left(A \cap \xi_j\right) \leq \mathbb{P}\left(\xi_j^c\right) + \mathbb{P}\left(A \cap \xi_j\right).$$

Therefore, applying recursively the previous inequality to bound the probability of error of P1 denoted $e_{\text{STO}}(n)$, we write

$$e_{\text{STO}}(n) = \mathbb{P}(\xi_2^c) \leq \sum_{i=2}^{K} \mathbb{P}\left(\xi_i^c \cap \left(\bigcap_{j=i+1}^{K+1} \xi_j\right)\right) \leq \sum_{i=2}^{K} \mathbb{P}\left(\xi_i^c \,\middle|\, \bigcap_{j=i+1}^{K+1} \xi_j\right).$$

Using Lemma 12; we get our claimed result in the stochastic case.

The adversarial case Given the adversary gain vector $\boldsymbol{g}$, the random variables $\widetilde{\boldsymbol{g}}_{k,t}$ can be dependent of each other for all $k \in [K]$ and $t \in [n]$ as $\boldsymbol{p}_{k,t}$ depends on previous observations at previous rounds. Therefore, we use the Bernstein inequality for martingale differences by Freedman (1975).

For random variables $\widetilde{\boldsymbol{g}}_{k,1}, \ldots, \widetilde{\boldsymbol{g}}_{k,n}$, we know that their variance is the variance of the Bernoulli random variable with parameter $1/\boldsymbol{p}_{k,t}$, scaled to the range $[0, \boldsymbol{g}_{k,t}/\boldsymbol{p}_{k,t}]$. For all $k \in [K]$ and $t \in [n]$, having a lower bound on $\boldsymbol{p}_{k,t} \geq 1/(K\overline{\log}K)$, we have $|\widetilde{\boldsymbol{g}}_{k,t} - \boldsymbol{g}_{k,t}| \leq K\overline{\log}K$ and

$$\sigma^2_{\widetilde{\boldsymbol{g}}_{k,t} - \boldsymbol{g}_{k,t}} = \sigma^2_{\widetilde{\boldsymbol{g}}_{k,t}} = \frac{\boldsymbol{p}_{k,t}(1 - \boldsymbol{p}_{k,t})\boldsymbol{g}_{k,t}^2}{\boldsymbol{p}_{k,t}^2} \leq K\overline{\log}K.$$

Then, following the same reasoning as in the proof of Theorem 1, but replacing the Bernstein inequality by the Bernstein inequality for martingale differences of Freedman (1975) applied to the martingale differences $\widetilde{\boldsymbol{g}}_{k,t} - \boldsymbol{g}_{k,t}$, we obtain the claimed result for the adversarial case. ■

Appendix G. On the complexities of $H_{\texttt{P1}}$, $H_{\texttt{SR}}$, and H_{BOB}

We now show that in general, $H_{\texttt{P1}} = \mathcal{O}\left(H_{\text{BOB}} \log^2 K\right)$. This demonstrates that P1 achieves the best that can be wished for in the two worlds, up to log factors. The extra $\log K$ from the $\log^2 K$ is not always present and we report an even more detailed discussion on the three different regimes of the gaps used in Remark 5 at the end of the section.

Corollary 7 *The complexity $H_{\texttt{P1}}$ of P1 matches the complexity H_{BOB} from the lower bound of Theorem 4 of up to log factors,*

$$H_{\texttt{P1}} = \mathcal{O}\left(H_{\text{BOB}} \log^2 K\right).$$

Proof To simplify the exposition, we assume, without loss of generality, that $na_i \in \mathbb{N}$, $\forall i \in [K]$. We set $j \triangleq \operatorname{argmin}(\Delta_k/k)$. Let $a_k \triangleq \Delta_{(1)}/\Delta_{(k)}, \forall k \in [K]$ and remember that $a_{K+1} = 0$. First note that the second term in $H_{\texttt{P1}}(\boldsymbol{a})$, taken for a fixed k, is of order (not considering the numerical constants and the $\overline{\log} K$)

$$\frac{K a_{\langle k \rangle} \Delta_k}{a_{\langle k \rangle}^2 \Delta_k^2} = \frac{K}{\Delta_{(1)}} \leq \frac{K}{\Delta_{(K)}\Delta_{(1)}} \leq \frac{j}{\Delta_j \Delta_{(1)}} = H_{\text{BOB}}.$$

Similarly, we have the first term in $H_{\texttt{P1}}(\boldsymbol{a})$ of order

$$\begin{aligned}
\frac{\sum_{i=\langle k \rangle}^{K} (a_i - a_{i+1}) i}{a_{\langle k \rangle}^2 \Delta_k^2} &= \frac{\sum_{i=\langle k \rangle}^{K-1} \left(\frac{\Delta_{(1)}}{\Delta_{(i)}} - \frac{\Delta_{(1)}}{\Delta_{(i+1)}}\right) i + K \frac{\Delta_{(1)}}{\Delta_{(K)}}}{\Delta_{(1)}^2} \\
&= \frac{\sum_{i=\langle k \rangle}^{K-1} \left(\frac{1}{\Delta_{(i)}} - \frac{1}{\Delta_{(i+1)}}\right) i + K \frac{1}{\Delta_{(K)}}}{\Delta_{(1)}} \\
&\overset{\textbf{(a)}}{\leq} \frac{\sum_{i=\langle k \rangle}^{K-1} \left(\frac{j}{i\Delta_{(j)}} - \frac{j}{(i+1)\Delta_{(j)}}\right) i + K \frac{j}{K\Delta_{(j)}}}{\Delta_{(1)}} \\
&= \frac{j\left(\sum_{i=\langle k \rangle}^{K} \left(\frac{1}{i} - \frac{1}{i+1}\right) i + 1\right)}{\Delta_{(1)}\Delta_{(j)}} \\
&= \frac{j\left(\sum_{i=\langle k \rangle}^{K} \frac{1}{i+1} + 1\right)}{\Delta_{(1)}\Delta_{(j)}} \\
&\leq H_{\text{BOB}} \left(\log K + 1\right),
\end{aligned}$$

where **(a)** is because as $j \triangleq \operatorname{argmin}_{k \in [K]}(\Delta_{(k)}/k)$, for all $i \in [K]$, $1/\Delta_{(i)} \leq j/(i\Delta_{(j)})$. More precisely, to see **(a)**, we unfold the sum and notice that actually there are no negative signs anywhere therefore, the upper bound holds. ■

G.1. Relation between $H_{\texttt{P1}}$, $H_{\texttt{SR}}$, and H_{BOB} for different regimes of the gaps.

We now study the relation between $H_{\texttt{P1}}$, $H_{\texttt{SR}}$, and H_{BOB} for different regimes of the gaps. We use the same examples as the ones used in Remark 5. We assume without loss of generality that $na_i \in \mathbb{N}$, $\forall i \in [2:K]$. In these three regimes of interest we prove that at worst, $H_{\texttt{P1}} = \mathcal{O}(H_{\text{BOB}} \log^2 K)$. In the flat regime and the square-root gap regime, we have $H_{\texttt{P1}} = \mathcal{O}(H_{\text{BOB}} \log K)$.

▶ **Flat regime** All the gaps are equal, that is, $k \in [2:K]$, we have that $\Delta_k = \Delta_1$.
We choose $a_i = 1,\ \forall i \in [2:K]$. First note that in $H_{\texttt{P1}}(\boldsymbol{a})$ the term

$$\frac{K a_{\langle k \rangle} \Delta_k}{a_{\langle k \rangle}^2 \Delta_k^2} = \frac{K}{\Delta_{(1)}} \leq \frac{K}{\Delta_{(1)}^2} = H_{\texttt{SR}} = H_{\texttt{BOB}}.$$

Then we have the first term

$$\frac{\sum_{i=\langle k \rangle}^{K} (a_i - a_{i+1}) i}{a_{\langle k \rangle}^2 \Delta_k^2} \leq \frac{K}{\min_{k \in [K]} \Delta_k^2} = H_{\texttt{SR}} = H_{\texttt{BOB}}.$$

▶ **Super-linear gaps** Since $(2) \in \operatorname{argmin}_k (\Delta_k / k)$, we get that $(2) \in \operatorname{argmin}_k (\Delta_k^2 / k)$.
Let $a_i \triangleq 1/i,\ \forall i \in [2:K]$. The two terms verify

$$\frac{K a_{\langle k \rangle} \Delta_k}{a_{\langle k \rangle}^2 \Delta_k^2} = \frac{K}{a_{\langle k \rangle} \Delta_k} = \frac{Kk}{\Delta_k} = \frac{2K}{\Delta_{(1)}} \leq \frac{2K}{\Delta_{(K)} \Delta_{(1)}} = \frac{2}{\Delta_{(1)}^2} = H_{\texttt{SR}} = H_{\texttt{BOB}} \quad \text{and}$$

$$\begin{aligned}
\frac{\sum_{i=\langle k \rangle}^{K} (a_i - a_{i+1}) i}{a_{\langle k \rangle}^2 \Delta_k^2} &= \frac{\sum_{i=\langle k \rangle}^{K} \left(\frac{1}{i} - \frac{1}{i+1} \right) i}{a_{\langle k \rangle}^2 \Delta_k^2} \\
&\leq \frac{4 \sum_{i=1}^{K} \left(\frac{1}{i} - \frac{1}{i+1} \right) i}{\Delta_1^2} \\
&= \frac{4 \sum_{i=1}^{K} \left(\frac{1}{i} (i+1) \right) i}{\Delta_1^2} \\
&\leq \frac{4 \log K}{\Delta_1^2} \\
&= 2 H_{\texttt{SR}} \log K \\
&= 2 H_{\texttt{BOB}} \log K.
\end{aligned}$$

▶ **Square-root gaps** We have that $(2) \in \operatorname{argmin}_k (\Delta_k^2 / k)$, $\sqrt{K/2} \Delta_{(1)} = \Delta_{(K)}$, and also that $\sqrt{k/2} \Delta_{(1)} \geq \Delta_{(k)}$ for $k \in [3:K-1]$. Therefore, we have $K \in \operatorname{argmin}_{k \in [K]} (\Delta_{(k)}^2 / k)$ and $K \in \operatorname{argmin}_{k \in [K]} (\Delta_{(k)} / k)$.
Let $a_i = 1/\sqrt{i},\ \forall i \in [2:K]$. The two terms verify

$$\frac{K a_{\langle k \rangle} \Delta_k}{a_{\langle k \rangle}^2 \Delta_k^2} = \frac{K}{a_{\langle k \rangle} \Delta_k} = \frac{K \sqrt{k}}{\Delta_k} = \frac{K \sqrt{2}}{\Delta_{(1)}} \leq \frac{K \sqrt{2}}{\Delta_{(K)} \Delta_{(1)}} = \sqrt{2} H_{\texttt{SR}} \sqrt{K} = \sqrt{2} H_{\texttt{BOB}} \quad \text{and}$$

$$\begin{aligned}
\frac{\sum_{i=\langle k\rangle}^{K}(a_i - a_{i+1})i}{a_{\langle k\rangle}^2\Delta_k^2} &= \frac{\sum_{i=\langle k\rangle}^{K}\left(\frac{1}{\sqrt{i}} - \frac{1}{\sqrt{i+1}}\right)i}{a_{\langle k\rangle}^2\Delta_k^2} \\
&\leq \frac{2\sum_{i=1}^{K}\frac{\sqrt{i}}{\sqrt{i+1}}\left(\sqrt{i+1} - \sqrt{i}\right)}{\Delta_{(1)}^2} \\
&\leq \frac{2\sum_{i=1}^{K}(\sqrt{i+1} - \sqrt{i})}{\Delta_{(1)}^2} \\
&\leq \frac{2\sqrt{K+1}}{\Delta_{(1)}^2} \\
&= \frac{2\sqrt{K+1}}{\Delta_{(1)}^2} \\
&= H_{\mathtt{SR}}\sqrt{K+1} \leq 2H_{\mathrm{BOB}}.
\end{aligned}$$