Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Minimax strategy for Stratified Sampling for Monte Carlo

Alexandra Carpentier ALEXANDRA.CARPENTIER@INRIA.FR
INRIA Lille - Nord Europe
40, avenue Halley
59650 Villeneuve d'Ascq , France

Remi Munos REMI.MUNOS@INRIA.FR
INRIA Lille - Nord Europe
40, avenue Halley
59650 Villeneuve d'Ascq , France

Andras Antos ANTOS@CS.BME.HU
Machine Learning Group
13-17 Kende u.,
H-1111 Budapest, Hungary

Editor: Unknown

Abstract

We consider the problem of stratified sampling for Monte-Carlo integration. We model this problem in a multi-armed bandit setting, where the arms represent the strata, and the goal is to estimate a weighted average of the mean values of the arms. We propose a strategy that samples the arms according to an upper bound on their standard deviations and compare its estimation quality to an ideal allocation that would know the standard deviations of the strata. We provide two pseudo-regret[1] analyses: a distribution-dependent bound of order $\widetilde{O}(n^{-3/2})$ that depends on a measure of the disparity of the strata, and a distribution-free bound $\widetilde{O}(n^{-4/3})$ that does not[2]. We also provide the first problem independent (minimax) lower bound for this problem and demonstrate that MC-UCB matches this lower bound both in terms of number of samples n and in terms of number of strata K. Finally, we link the pseudo-regret with the difference between the mean squared error on the estimated weighted average of the mean values of the arms, and the optimal "oracle" strategy: this provides us also a problem dependent and a problem independent rate for this measure of performance and, as a corollary, asymptotic optimality.

Keywords: Bandit Theory, Stratified Monte-Carlo, Minimax strategies.

1. Introduction

Consider a polling institute that has to estimate as accurately as possible the average income of a country, given a finite budget for polls. The institute has call centers in every region in the country, and gives a part of the total sampling budget to each center so that they can call random people in the area and ask about their income. A naive method would allocate a budget proportionally to the number of people in each area. However some regions show

1. We define this notion in Section 2. It is a proxy on the difference between the mean squared error on the estimated weighted average of the mean values of the arms, and the optimal "oracle" strategy.
2. The notation $\widetilde{O}(\cdot)$ corresponds to $O(\cdot)$ up to logarithmic factors.

a high variability in the income of their inhabitants whereas others are very homogeneous. Now if the polling institute knows the level of variability within each region, it could adjust the budget allocated to each region in a more clever way (allocating more polls to regions with high variability) in order to reduce the final estimation error.

This example is just one of many for which an efficient method of sampling a function with natural strata (i.e., the regions) is of great interest. Note that even in the case that there are no natural strata, it is always a good strategy to design arbitrary strata and allocate a budget to each stratum that is proportional to the size of the stratum, compared to a crude Monte-Carlo. There are many good surveys on the topic of stratified sampling for Monte-Carlo, such as (Rubinstein and Kroese, 2008)[Subsection 5.5] or (Glasserman, 2004).

The main problem for performing an efficient sampling is that the variances within the strata (in the previous example, the income variability per region) are unknown. One possibility is to estimate the variances *online* while sampling the strata. There is some interesting research along this direction, such as (Arouna, 2004) and more recently (Etoré and Jourdain, 2010; Kawai, 2010). The work of Etoré and Jourdain (2010) matches exactly our problem of designing an efficient adaptive sampling strategy. In this article they propose to sample according to an empirical estimate of the variance of the strata, whereas Kawai (2010) addresses a computational complexity problem which is slightly different from ours. The recent work of Etoré et al. (2011) describes a strategy that enables to sample *asymptotically* according to the (unknown) standard deviations of the strata and at the same time adapts the shape (and number) of the strata online. This is a very difficult problem, especially in high dimension, that we will not address here, although we think this is a very interesting and promising direction for further researches.

These works provide asymptotic convergence of the variance of the estimate to the targeted stratified variance [3] divided by the sample size. They also prove that the number of pulls within each stratum converges asymptotically to the desired number of pulls i.e. the optimal allocation if the variances per stratum were known. Like Etoré and Jourdain (2010), we consider a stratified Monte-Carlo setting with fixed strata. Our contribution is to design a sampling strategy for which we can derive a finite-time analysis (where 'time' refers to the number of samples). This enables us to predict the quality of our estimate for any given budget n.

We model this problem using the setting of multi-armed bandits where our goal is to estimate a weighted average of the mean values of the arms. Although our goal is different from a usual bandit problem where the objective is to play the best arm as often as possible, this problem also exhibits an *exploration-exploitation trade-off*. The arms have to be pulled both in order to estimate the initially unknown variability of the arms (exploration) and to allocate correctly the budget according to our current knowledge of the variability (exploitation).

Our setting is close to the one described in (Antos et al., 2010) which aims at estimating *uniformly well* the mean values of all the arms. The authors present an algorithm, called GAFS-MAX, that allocates samples proportionally to the empirical variance of the arms, while imposing that each arm is pulled at least $\sqrt{n}$ times to guarantee a sufficiently good

3. The target is defined in [Subsection 5.5] of (Rubinstein and Kroese, 2008) and later in this paper, see Equation 4.

estimation of the true variances. Another approach for this problem, still with a bandit formalism, can be found in (Carpentier et al., 2011), and the analysis is extended.

Note tough that in the Master Thesis (Grover, 2009), the author presents an algorithm named GAFS-WL which is similar to GAFS-MAX and has an analysis close to the one of GAFS-MAX. It deals with stratified sampling, i.e. it targets an allocation which is proportional to the standard deviation (and not to the variance) of the strata time their size[4]. They define a proxy on the mean squared error that they write *loss*, and prove that the difference between the loss of GAFS-WL and the optimal static loss is of order $\tilde{O}(n^{-3/2})$, where the $\tilde{O}(.)$ depends of the problem. There are however some open questions in this very good Master Thesis. A first one is on the existence of a problem dependent bound for GAFS-WL. A second important issue is on the links between the loss they define and the intuitive, related measure of performance, which is the mean squared error. Without this link, they are not able to prove that GAFS-WL is asymptotically optimal.

Our objective is similar, and we extend the analysis of this setting. We introduced in paper (Carpentier and Munos, 2011) algorithm MC-UCB, a new algorithm based on Upper-Confidence-Bounds (UCB) on the standard deviations. They are computed from the empirical standard deviation and a confidence interval derived from Bernstein's inequalities. The algorithm, called MC-UCB, samples the arms proportionally to an UCB[5] on the standard deviation times the size of the stratum. We provided finite-time, problem dependent and problem independent bounds for the loss of this algorithm, filling the gap in (Grover, 2009). We however, as in (Grover, 2009), did not link this pseudo-regret to the mean squared-error.

Contributions: In this paper we extend the analysis of MC-UCB in (Carpentier and Munos, 2011). Our contributions are the following:

- We detail the proofs of paper (Carpentier and Munos, 2011), which have not been published in this version due to space constraints. They correspond to two pseudo-regret analysis: (i) a distribution-dependent bound of order $\widetilde{O}(n^{-3/2})$ that depends on the disparity of the stratas (a measure of the problem complexity), and which corresponds to a stationary regime where the budget n is large compared to this complexity. (ii) A distribution-free bound of order $\widetilde{O}(n^{-4/3})$ that does not depend on the the disparity of the stratas, and corresponds to a transitory regime where n is small compared to the complexity. The characterization of those two regimes and the fact that the corresponding excess error rates differ enlightens the fact that a finite-time analysis is very relevant for this problem.

- More precisely, we improve the problem independent upper bound in terms of K. This bound on the expectation of the pseudo-regret is of order $\tilde{O}(\frac{K^{1/3}}{n^{4/3}})$ where K is the number of strata.

- We also provide a minimax lower bound on the expectation of the pseudo-regret for the problem of stratified Monte-Carlo of order $\Omega(\frac{K^{1/3}}{n^{4/3}})$. As a matter of fact, the

4. This is explained in (Rubinstein and Kroese, 2008) and will be formulated precisely later.
5. Note that we consider a sampling strategy based on UCBs on the standard deviations of the arms whereas the so-called *UCB algorithm* of Auer et al. (2002), in the usual multi-armed bandit setting, computes UCBs on the mean rewards of the arms.

problem independent lower-bound matches the problem independent upper-bound for MC-UCB, in terms of n and K. It induces that MC-UCB is minimax optimal in terms of pseudo-regret.

- Finally, by clarifying the notion of pseudo-regret that we introduce in Section 2, we provide finite-time bound on the mean squared error of the estimate of the integral. As a corollary, we obtain also asymptotic consistency of our algorithm.

The rest of the paper is organized as follows. In Section 2 we formalize the problem and introduce the notations used throughout the paper. Section 3 states the minimax lower bound on the pseudo-regret. Section 4 introduces the MC-UCB algorithm and reports performance bounds. Section 5 discusses the bridges between the pseudo regret and the mean squared error. We then discuss in Section 6 about the parameters of the algorithm and its performances. In Section 7 we report numerical experiments that illustrate our method to the problem of pricing Asian options as introduced in (Glasserman et al., 1999). Finally, Section 8 concludes the paper and suggests future works.

2. Preliminaries

The allocation problem mentioned in the previous section is formalized as a K-armed bandit problem where each arm (stratum) $k = 1, \dots, K$ is characterized by a distribution ν_k with mean value μ_k and variance σ_k^2. At each round $t \geq 1$, an allocation strategy (or algorithm) $\mathcal{A}$ selects an arm k_t and receives a sample drawn from ν_{k_t} independently of the past samples. Note that a strategy may be adaptive, i.e., the arm selected at round t may depend on past observed samples. Let $\{w_k\}_{k=1,\dots,K}$ denote a known set of positive weights which sum to 1. For example in the setting of stratified sampling for Monte-Carlo, this would be the probability mass in each stratum. The goal is to define a strategy that estimates as precisely as possible $\mu = \sum_{k=1}^K w_k \mu_k$ using a total budget of n samples.

Let us write $T_{k,t} = \sum_{s=1}^t \mathbb{I}\{k_s = k\}$ the number of times arm k has been pulled up to time t, and $\hat{\mu}_{k,t} = \dfrac{1}{T_{k,t}} \sum_{s=1}^{T_{k,t}} X_{k,s}$ the empirical estimate of the mean μ_k at time t, where $X_{k,s}$ denotes the sample received when pulling arm k for the s-th time.

After n rounds, the algorithm $\mathcal{A}$ returns the empirical estimate $\hat{\mu}_{k,n}$ of all the arms. Note that in the case of a deterministic strategy, the expected quadratic estimation error of the weighted mean μ as estimated by the weighted average $\hat{\mu}_n = \sum_{k=1}^K w_k \hat{\mu}_{k,n}$ satisfies:

$$\mathbb{E}\Big[\big(\hat{\mu}_n - \mu\big)^2\Big] = \mathbb{E}\Big[\big(\textstyle\sum_{k=1}^K w_k(\hat{\mu}_{k,n} - \mu_k)\big)^2\Big] = \textstyle\sum_{k=1}^K w_k^2 \frac{\sigma_k^2}{T_{k,n}},$$

where $\mathbb{E}\Big[.\Big]$ is the expectation integrated over all the samples of all arms.

We thus use the following measure for the performance of any algorithm $\mathcal{A}$:

$$L_n(\mathcal{A}) = \textstyle\sum_{k=1}^K w_k^2 \frac{\sigma_k^2}{T_{k,n}}. \tag{1}$$

We denote this quantity by pseudo-loss, as it is a proxy of the true loss of the algorithm, which is $\mathbb{E}\Big[\big(\hat{\mu}_n - \mu\big)^2\Big]$. This loss is not the same as in (Grover, 2009) and in (Carpentier and

Munos, 2011). We give some properties of this pseudo-loss in Section 5. We also provide in Subsection 5.1 properties of the loss defined in papers (Grover, 2009) and (Carpentier and Munos, 2011).

The goal is to define an allocation strategy that minimizes the global pseudo-loss defined in Equation 1. If the variance of the arms were known in advance, one could design an optimal static[6] allocation strategy $\mathcal{A}^*$ by pulling each arm k proportionally to the quantity $w_k\sigma_k$. Indeed, if arm k is pulled a deterministic number of times $T^*_{k,n}$, then [7]

$$L_n(\mathcal{A}^*) = \sum_{k=1}^K w_k^2 \frac{\sigma_k^2}{T^*_{k,n}} \ . \tag{2}$$

By choosing $T^*_{k,n}$ such as to minimize L_n under the constraint that $\sum_{k=1}^K T^*_{k,n} = n$, the optimal static allocation (up to rounding effects) of algorithm $\mathcal{A}^*$ is to pull each arm k,

$$T^*_{k,n} = \frac{w_k\sigma_k}{\sum_{i=1}^K w_i\sigma_i} n \ , \tag{3}$$

times, and achieves a global pseudo-loss (or loss as the $(T^*_{k,n})_k$ are deterministic)

$$L_n(\mathcal{A}^*) = \frac{\Sigma_w^2}{n} \ , \tag{4}$$

where $\Sigma_w = \sum_{i=1}^K w_i\sigma_i$ (we assume in the sequel that $\Sigma_w > O$). In the following, we write $\lambda_k = \frac{T^*_{k,n}}{n} = \frac{w_k\sigma_k}{\Sigma_w}$ the optimal allocation proportion for arm k and $\lambda_{\min} = \min_{1\le k\le K} \lambda_k$. Note that a small $\lambda_{\min}$ means a large disparity of the $w_k\sigma_k$ and, as explained later, provides for the algorithm we build in Section 4 a characterization of the hardness of a problem.

However, in the setting considered here, the σ_k are unknown, and thus the optimal allocation is out of reach. A possible allocation is the uniform strategy $\mathcal{A}^u$, i.e., such that $T^u_k = \frac{w_k}{\sum_{i=1}^K w_i} n$. Its pseudo-loss (and loss as the $(T^u_k)_k$ are deterministic) is

$$L_n(\mathcal{A}^u) = \sum_{k=1}^K w_k \sum_{k=1}^K \frac{w_k\sigma_k^2}{n} = \frac{\Sigma_{w,2}}{n} \ ,$$

where $\Sigma_{w,2} = \sum_{k=1}^K w_k\sigma_k^2$. Note that by Cauchy-Schwartz's inequality, we have $\Sigma_w^2 \le \Sigma_{w,2}$ with equality if and only if the $(\sigma_k)_k$ are all equal. Thus $\mathcal{A}^*$ is always at least as good as $\mathcal{A}^u$. In addition, since $\sum_i w_i = 1$, we have $\Sigma_w^2 - \Sigma_{w,2} = -\sum_k w_k(\sigma_k - \Sigma_w)^2$. The difference between those two quantities is the weighted quadratic variation of the σ_k around their weighted mean Σ_w. In other words, it is the variance of the $(\sigma_k)_{1\le k\le K}$. As a result the gain of $\mathcal{A}^*$ compared to $\mathcal{A}^u$ grow with the disparity of the σ_k.

We would like to do better than the uniform strategy by considering an adaptive strategy $\mathcal{A}$ that would estimate the σ_k at the same time as it tries to implement an allocation strategy as close as possible to the optimal allocation algorithm $\mathcal{A}^*$. This introduces a natural trade-off between the exploration needed to improve the estimates of the variances and the exploitation of the current estimates to allocate the pulls nearly-optimally.

6. Static means that the number of pulls allocated to each arm does not depend on the received samples.
7. As it will be discussed later, this equality does not hold when the number of pulls is random, as it is the case of adaptive algorithms where the strategy depends on the observed samples.

In order to assess how well $\mathcal{A}$ solves this trade-off and manages to sample according to the true standard deviations *without knowing them in advance*, we compare its performance to that of the optimal allocation strategy $\mathcal{A}^*$. For this purpose we define the notion of *pseudo-regret* of an adaptive algorithm $\mathcal{A}$ as the difference between the pseudo-loss incurred by the algorithm and the optimal pseudo-loss:

$$R_n(\mathcal{A}) = L_n(\mathcal{A}) - L_n(\mathcal{A}^*). \tag{5}$$

The *pseudo-regret* indicates how much we loose in terms of expected quadratic estimation error by not knowing in advance the standard deviations (σ_k). Note that since $L_n(\mathcal{A}^*) = \frac{\Sigma_w^2}{n}$, a consistent strategy i.e., asymptotically equivalent to the optimal strategy, is obtained whenever its regret is negligible compared to $1/n$.

We also defined the *true regret* as

$$\bar{R}_n(\mathcal{A}) = \mathbb{E}[(\hat{\mu}_n - \mu)^2] - L_n(\mathcal{A}^*). \tag{6}$$

This is the difference between the mean-squared error and the optimal mean squared error. The pseudo-regret is a proxy for the true regret.

3. Minimax lower-bound on the pseudo-regret

We now study the minimax rate for the pseudo-regret of any algorithm on a given stratification in K strata of equal size.

Theorem 1 *Let* inf *be the infimum taken over all online stratified sampling algorithms using K strata and* sup *represent the supremum taken over all environments, then:*

$$\inf \sup \mathbb{E} R_n \geq C \frac{K^{1/3}}{n^{4/3}},$$

where C is a numerical constant.

Proof [Sketch of proof (The full proof is reported in Appendix A)] We consider a stratification with $2K$ strata. On the K first strata, the samples are drawn from Bernoulli distributions of parameter μ_k where $\mu_k \in \{\frac{\mu}{2}, \mu, 3\frac{\mu}{2}\}$, and on the K last strata, the samples are drawn from a Bernoulli of parameter $1/2$. We write $\sigma = \sqrt{\mu(1-\mu)}$ the standard deviation of a Bernoulli of parameter μ. We index by ϵ a set of 2^K possible environments, where $\epsilon = (\epsilon_1, \ldots, \epsilon_K) \in \{-1, +1\}^K$, and the K first strata are defined by $\mu_k = \mu + \epsilon_k \frac{\mu}{2}$. Write $\mathbb{P}_\sigma$ the probability under such an environment, also consider $\mathbb{P}_\sigma$ the probability under which all the K first strata are Bernoulli with mean μ.

We define Ω_ϵ the event on which there are less than $\frac{K}{3}$ arms not pulled correctly for environment ϵ (i.e. for which $T_{k,n}$ is larger than the optimal allocation corresponding to μ when actually $\mu_k = \frac{\mu}{2}$, or smaller than the optimal allocation corresponding to μ when $\mu_k = 3\frac{\mu}{2}$). See the Appendix A for a precise definition of these events. Then, the idea is that there are so many such environments that any algorithm will be such that for at least one of them we have $\mathbb{P}_\sigma(\Omega_\epsilon) \leq \exp(-K/72)$. Then we derive by a variant of Pinsker's inequality applied to an event of small probability that $\mathbb{P}_\epsilon(\Omega_\epsilon) \leq \frac{KL(\mathbb{P}_\sigma, \mathbb{P}_\epsilon)}{K} = O(\frac{\sigma^{3/2} n}{K})$.

Finally, by choosing σ of order $(\frac{K}{n})^{1/3}$, we have that $\mathbb{P}_\epsilon(\Omega_\epsilon^c)$ is bigger than a constant, and on Ω_ϵ^c we know that there are more than $\frac{K}{3}$ arms not pulled correctly. This leads to an expected pseudo-regret in environment ϵ of order $\Omega(\frac{K^{1/3}}{n^{4/3}})$. ■

This is the first lower-bound for the problem of online stratified sampling for Monte-Carlo. We sketch the proof in the main text because we believe that the technique of proof for this bound is original. It follows from the fact that no algorithm can allocate the samples in *every* problem according to the unknown best proportions with a better precision than $\frac{n^{2/3}}{K^{2/3}}$ for a number of arms non negligible when compared to K, with a probability larger than a non negligible constant.

4. Allocation based on Monte Carlo Upper Confidence Bound

4.1 The algorithm

In this section, we introduce our adaptive algorithm for the allocation problem, called *Monte Carlo Upper Confidence Bound* (MC-UCB). The algorithm computes a high-probability bound on the standard deviation of each arm and samples the arms proportionally to their bounds times the corresponding weights. The MC-UCB algorithm, $\mathcal{A}_{MC-UCB}$, is described in Figure 1. It requires three parameters as inputs: c_1 and c_2 which are related to the shape of the distributions (see Assumption 1), and δ which defines the *confidence level* of the bound. In Subsection 6.4, we discuss a way to reduce the number of parameters from three to one. The amount of exploration of the algorithm can be adapted by properly tuning these parameters.

Input: c_1, c_2, δ. Let $a = \sqrt{2\log(2/\delta)}\sqrt{c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+\log(c_2/\delta))}n^{1/2}}{2(1-\delta)}$.
Initialize: Pull each arm twice.
for $t = 2K+1, \ldots, n$ **do**
  Compute $B_{k,t} = \frac{w_k}{T_{k,t-1}}\left(\hat{\sigma}_{k,t-1} + 2a\sqrt{\frac{1}{T_{k,t-1}}}\right)$ for each arm $1 \le k \le K$
  Pull an arm $k_t \in \arg\max_{1\le k\le K} B_{k,t}$
end for
Output: $\hat{\mu}_{k,t}$ for each arm $1 \le k \le K$

Figure 1: The pseudo-code of the MC-UCB algorithm. The empirical standard deviations $\hat{\sigma}_{k,t-1}$ are computed using Equation 7.

The algorithm starts by pulling each arm twice in rounds $t = 1$ to $2K$. From round $t = 2K+1$ on, it computes an upper confidence bound $B_{k,t}$ on the standard deviation σ_k, for each arm k, and then pulls the one with largest $B_{k,t}$. The upper bounds on the standard deviations are built by using Theorem 10 in (Maurer and Pontil, 2009)[8] and based on the

8. We could also have used the variant reported in (Audibert et al., 2009).

empirical standard deviation $\hat{\sigma}_{k,t-1}$:

$$\hat{\sigma}^2_{k,t-1} = \frac{1}{T_{k,t-1}-1} \sum_{i=1}^{T_{k,t-1}} (X_{k,i} - \hat{\mu}_{k,t-1})^2, \tag{7}$$

where $X_{k,i}$ is the i-th sample received when pulling arm k, and $T_{k,t-1}$ is the number of pulls allocated to arm k up to time $t-1$. After n rounds, MC-UCB returns the empirical mean $\hat{\mu}_{k,n}$ for each arm $1 \leq k \leq K$.

4.2 Pseudo-Regret analysis of MC-UCB

Before stating the main results of this section, we state the assumption that the distributions are sub-Gaussian, which includes e.g., Gaussian or bounded distributions. See (Buldygin and Kozachenko, 1980) for more precisions.

Assumption 1 *There exist $c_1, c_2 > 0$ such that for all $1 \leq k \leq K$ and any $\epsilon > 0$,*

$$\mathbb{P}_{X \sim \nu_k}(|X - \mu_k| \geq \epsilon) \leq c_2 \exp(-\epsilon^2/c_1) \ . \tag{8}$$

We provide two analyses, a *distribution-dependent* and a *distribution-free*, of MC-UCB, which are respectively interesting in two *regimes*, i.e., stationary and transitory *regimes*, of the algorithm. We will comment on this later in Section 6.

A *distribution-dependent* result: We now report the first bound on the expectation of the pseudo-regret of MC-UCB algorithm. The proof is reported in Appendix C (in the supplementary material) and relies on upper- and lower-bounds on $T_{k,t} - T^*_{k,t}$, i.e., the difference in the number of pulls of each arm compared to the optimal allocation (see Lemma 15).

Theorem 2 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, the pseudo-regret of MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$ is bounded in expectation as*

$$\mathbb{E}[R_n] \leq 336\sqrt{2c_1(c_2+2)}(\sqrt{c_2}+1)^{2/3} K^{1/3} \Sigma_w \frac{\log(n)}{n^{4/3}} + \frac{5K\Sigma_{w,2}}{n^2}.$$

Note that this result crucially depends on the smallest proportion $\lambda_{\min}$ which is a measure of the disparity of product of the standard deviations and the weights. For this reason we refer to it as "distribution-dependent" result. The full proof for this result is in Appendix C.

A *distribution-free* result: Now we report our second pseudo-regret bound that does not depend on $\lambda_{\min}$ but whose rate is poorer. The proof is given in Appendix D of the supplementary material and relies on other upper- and lower-bounds on $T_{k,t} - T^*_{k,t}$ detailed in Lemma 16.

Theorem 3 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, the pseudo-regret of MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$ is bounded in expectation as*

$$\mathbb{E}[R_n] \leq \frac{\Sigma_w^2}{n} + 336\sqrt{2c_1(c_2+2)}(\sqrt{c_2}+1)^{2/3} K^{1/3} \Sigma_w \frac{\log(n)}{n^{4/3}} + \frac{5K\Sigma_{w,2}}{n^2}.$$

This bound does not depend on $1/\lambda_{\min}$, not even in the negligible term, as detailed in Appendix D[9]. This is obtained at the price of the slightly worse rate $\widetilde{O}(n^{-4/3})$.

5. Links between the pseudo-loss and the mean-squared error

As mentioned in Section 2, the pseudo-loss is trivially equal to the mean-squared error of the estimate $\hat{\mu}_n$ of μ if the number of samples $T_{k,n}$ in each stratum is independent of the samples. This is not the case for any reasonable adaptive strategy, as such methods precisely aim at adapting the number of samples in each stratum to the standard deviation inside the stratum.
It is however important to derive links between those two quantities, in order for the pseudo-loss and the pseudo-regret to be meaningful. The mean squared error can be decomposed as

$$\mathbb{E}\big[(\hat{\mu}_n - \mu)^2\big] = \sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big] + \sum_{k=1}^{n} \sum_{k' \neq k} w_k w_q \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\big].$$

The quantity $\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big]$ is equal to the loss defined in (Grover, 2009) and (Carpentier and Munos, 2011). If the $(T_{k,n})_k$ are deterministic, this quantity is equal to the pseudo-loss and also to the mean squared error $\mathbb{E}\big[(\hat{\mu}_n - \mu)^2\big]$. If the $(T_{k,n})_k$ are deterministic, the cross-products $\sum_{k=1}^{n} \sum_{k' \neq k} w_k w_q \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\big]$ are equal to 0.

A natural way to proceed is to (i) prove that the expectation of the pseudo-loss is not very different from $\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big]$ (and thus from $\frac{\Sigma_w^2}{n}$) and (ii) prove that the cross-products are close to 0.

5.1 Bounds on $\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big]$

The technique for bounding $\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big]$ is very similar to the one for bounding the expectation of the pseudo-loss. The only additional technical passage is to use Wald's identity to bound $\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big]$ with a quantity close to the expectation of the pseudo-loss.

We have in the same way a problem dependent bound and a problem independent bound.

Problem dependent bound.

Proposition 4 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, then for algorithm MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$, we have*

$$\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big] - \frac{\Sigma_w^2}{n}$$
$$\leq \frac{\log(n)}{n^{3/2}\lambda_{\min}^{3/2}} \Big(112\Sigma_w \sqrt{c_1(c_2+2)} + 6c_1(c_2+2)K\Big) + \frac{19}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 720c_1(c_2+1)\log(n)^2\Big).$$

The full proof is in Appendix C.

9. Note that the bound is not entirely distribution free since Σ_w appears. But it can be proved using Assumption 1 that $\Sigma_w^2 \leq c_1 c_2$.

Problem independent bound.

Proposition 5 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, then for algorithm MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$, we have*

$$\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big] - \frac{\Sigma_w^2}{n}$$
$$\leq \frac{200\sqrt{c_1}(c_2+2)\Sigma_w K}{n^{4/3}} \log(n) + \frac{365}{n^{3/2}}\Big(129 c_1 (c_2+2)^2 K^2 \log(n)^2 + K\Sigma_w^2\Big).$$

The full proof is in Appendix D.

5.2 Bounds on the cross-products

The difficulty in bounding the cross-product comes from the fact that the $(T_{k,n})_k$ depend on the samples, and more exactly for algorithm MC-UCB, on the sequence of empirical standard deviations $(\sigma_{k,t})_{t\leq n}$ of each arm k. As in general $\hat{\mu}_{k,n}$ depends on $(\sigma_{k,t})_{t\leq n}$, there is no direct reason why the cross-products should be equal to 0.

We prove three results for bounding these cross-products. The first one corresponds to the specific case where the distribution of the arms are symmetric. We then provide a problem dependent and a problem independent bound in the general case.

Equality holds when the distributions of the arms are symmetric. A first result is in the specific case of symmetric distributions. Intuitively in this setting, the empirical standard deviations are independent of the signs of $(\hat{\mu}_{k,n} - \mu_k)$. This implies that the signs of $(\hat{\mu}_{k,n} - \mu_k)$ and $(\hat{\mu}_{q,n} - \mu_q)$ are independent of each other when $k \neq q$. From that we deduce the following result.

Proposition 6 *Assume that the distributions $(\nu_k)_k$ of the arms are symmetric around μ_k respectively. For algorithm MC-UCB launched with any parameters, we have*

$$\sum_{k=1}^{n} \sum_{k' \neq k} w_k w_q \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\big] = 0.$$

The proof of this result is to be found in Appendix F.1.

Problem dependent bound in the general case. On an event of high probability, $|T_{k,n} - T^*_{k,n}| = \tilde{O}(n^{-1/2})$ as explained in Lemma 15 in the Appendices[10]. This means that even though $T_{k,n}$ is random, it does not deviate too much from $T^*_{k,n}$. From that we deduce the following problem dependent bound.

Proposition 7 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, then for algorithm MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$, we have*

$$\sum_{k=1}^{n} \sum_{k' \neq k} w_k w_q \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\big] \leq \tilde{O}(n^{-3/2}),$$

where $\tilde{O}(.)$ hides an invert dependency in $\lambda_{\min}$.

The proof of this result is in Appendix F.2

10. Here $\tilde{O}(\cdot)$ depends on $\lambda_{\min}^{-1}$.

Problem independent bound in the general case. On an event of high probability, $|T_{k,n} - T^*_{k,n}| = \tilde{O}(n^{-2/3})$ as explained in Lemma 16 in the Appendices. From that we deduce in the same way that for he previous proposition the following problem independent bound.

Proposition 8 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, then for algorithm MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$, we have*

$$\sum_{k=1}^{n} \sum_{k' \neq k} w_k w_q \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\big] \leq \tilde{O}(n^{-7/6}),$$

where $\tilde{O}(.)$ does not depend on $\lambda_{\min}$.

The proof of this result is in Appendix F.2.

5.3 Bounds on the true regret and asymptotic optimality

We are finally able to fulfill the objective of this Section, that is to say bound the true regret $\bar{R}_n = \mathbb{E}[(\hat{\mu}_n - \mu)^2] - \frac{\Sigma_w^2}{n}$. We have the following Theorem directly by combining the results of the Propositions in Subsections 5.1 and 5.

Theorem 9 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, then for algorithm MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$, the true regret is bounded as*

$$\bar{R}_n = \tilde{O}(n^{-3/2}),$$

where $\tilde{O}(.)$ hides a dependency in $\lambda_{\min}^{-1}$, and

$$\bar{R}_n = \tilde{O}(n^{-7/6}),$$

where $\tilde{O}(.)$ does not depend on $\lambda_{\min}$.

An immediate corollary on asymptotic optimality follows, when the parameter δ_n (for a given budget n) is chosen wisely.

Corollary 10 *Under Assumption 1 and if we choose c_2 such that $c_2 \geq 2Kn^{-5/2}$, then for algorithm MC-UCB launched with parameter $\delta = n^{-7/2}$ with $n \geq 4K$, the true regret converges and*

$$\lim_{n \to +\infty} \bar{R}_n = 0.$$

Proof [Proof of Corollary 10] The proof follows directly from Borel-Cantelli, as $\sum_n \delta_n < +\infty$. ■

6. Discussion on the results

We make several comments on the algorithm $MC - UCB$ in this Section.

6.1 Problem dependent and independent bounds for the expectation of the pseudo-loss

Theorem 2 provides a pseudo-regret bound of order $\widetilde{\lambda}_{\min}^{-3/2} O(n^{-3/2})$, whereas Theorem 3 provides a bound of order $\widetilde{O}(n^{-4/3})$ independently of $\lambda_{\min}$. Hence, for a given problem i.e., a given $\lambda_{\min}$, the distribution-free result of Theorem 3 is more informative than the distribution-dependent result of Theorem 2 in the *transitory regime*, that is to say when n is small compared to $\lambda_{\min}^{-1}$. The distribution-dependent result of Theorem 2 is better in the *stationary regime* i.e., for n large. This distinction reminds us of the difference between distribution-dependent and distribution-free bounds for the UCB algorithm in usual multi-armed bandits[11].

The problem dependent lower bound is similar to the one provided for GAFS-WL in (Grover, 2009). In their article, their pseudo-loss measure is $\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big]$ so we compare their bound with the ones in Propositions 4 and 5. We however expect that GAFS-WL has for some problems a sub-optimal behavior: it is possible to find cases where $\mathbb{E}\Big[\sum_k w_k^2 (\hat{\mu}_{k,n} - \mu_k)^2\Big] - \frac{\Sigma_w^2}{n} \geq O(1/n)$, see Appendix E for more details. It is not the case for MC-UCB, for which $\mathbb{E}\Big[\sum_k w_k^2 (\hat{\mu}_{k,n} - \mu_k)^2\Big] - \frac{\Sigma_w^2}{n} \leq \tilde{O}(n^{-4/3})$. Note however that when there is an arm with 0 standard deviation, GAFS-WL is likely to perform better than MC-UCB, as it will only sample this arm $O(\sqrt{n})$ times while MC-UCB samples it $\widetilde{O}(n^{2/3})$ times.

6.2 Finite-time bounds for $\mathbb{E}[(\hat{\mu}_n - \mu)^2]$, and on the true regret, and asymptotic optimality

We also bound the true regret $\bar{R}_n = \mathbb{E}[(\hat{\mu}_n - \mu)^2] - \frac{\Sigma_w^2}{n}$ in $o(\frac{1}{n})$. This means that the mean squared error of the estimate is very close to the "oracle" smallest mean squared error possible, obtained with a deterministic strategy that has access to $(\sigma_k)_k$.

The first result in Theorem 9 states that for MC-UCB, the true regret is of order $\tilde{O}(n^{-3/2})$, where the $\tilde{O}$ hides a dependency in $\lambda_{\min}$. This is the equivalent of the problem dependent bound on the pseudo-loss. This Theorem also states that for MC-UCB, an upper bound on the true regret is of order $\tilde{O}(n^{-7/6})$, where the $\tilde{O}$ does not depend in any way on $\lambda_{\min}$. This is the equivalent of the problem independent bound on the pseudo-loss. Unfortunately, we do not obtain a problem independent bound that is of the same order as the problem independent bound of the pseudo-regret, i.e. $\tilde{O}(n^{-4/3})$. This comes from the fact that the bound on the cross-products in Proposition 8 is of order $\tilde{O}(n^{-7/6})$. Whether this bound is tight or not is an open problem.

These results imply that algorithm MC-UCB is asymptotically optimal (like the algorithms of Kawai (2010); Etoré and Jourdain (2010)): the estimate $\hat{\mu}_n = \sum_k w_k \hat{\mu}_{k,n}$ is asymptotically equal to μ and the variance of $\hat{\mu}_n$ is asymptotically equal to the variance of the optimal allocation Σ_w^2/n for *any* problem. Note that the asymptotic optimality of GAFS-WL is not provided in Grover (2009), although we believe it to hold.

11. The distribution dependent bound is in $O(K \log n/\Delta)$, where Δ is the difference between the mean value of the two best arms, and the distribution-free bound is in $O(\sqrt{nK \log n})$ as explained in (Auer et al., 2002; Audibert and Bubeck, 2009).

Note also that whenever there is some disparity among the arms, i.e., when $\Sigma_w^2 - \Sigma_{2,w} < 0$, the MC-UCB is asymptotically strictly more efficient than the uniform strategy.

6.3 MC-UCB and the lower bound

We provide in this paper a minimax (problem independent) lower-bound for the pseudo-regret that is in expectation of order $\Omega(\frac{K^{1/3}}{n^{4/3}})$ (see Theorem 1). An important achievement is that the problem independent upper bound on the pseudo-regret of MC-UCB is in expectation of the same order up to a logarithmic factor (see Theorem 3). It is thus impossible to improve this strategy uniformly on every problem, more than by a log factor.

Although we do not have a problem dependent lower bound on the pseudo-regret yet, we believe that the rate $\tilde{O}(n^{-3/2})$ cannot be improved for general distributions. As explained in the proof in Appendix C, this rate is a direct consequence of the high probability bounds on the estimates of the standard deviations of the arms which are in $O(1/\sqrt{n})$, *and those bounds are tight*. Because of the minimax lower-bound that is of order $O(n^{-4/3})$, it is however clear that there exists no algorithm with a regret of order $\widetilde{O}(n^{-3/2})$ without any dependence in $\lambda_{\min}^{-1}$ (or another related problem-dependent quantity).

6.4 The parameters of the algorithm

Our algorithm takes three parameters as input, namely c_1, c_2 and δ, but we only use a combination of them in the algorithm, with the introduction of $a = \sqrt{2\log(2/\delta)}\sqrt{c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+\log(c_2/\delta))}n^{1/2}}{2(1-\delta)}$. For practical use of the method, it is enough to tune the algorithm with a single parameter a. By the choice of the value assigned to δ in the two theorems, $a \approx c\log(n)$, where c can be interpreted as a high probability bound on the range of the samples. We thus simply require a rough estimate of the magnitude of the samples. Note that in the case of bounded distributions, a can be chosen as $a = 2\sqrt{\frac{5}{2}}c\sqrt{\log(n)}$ where c is a true bound on the variables. This result is easy to deduce by simplifying Lemma 11 in Appendix B for the case of bounded variables.

6.5 Making MC-UCB anytime

An interesting question is on whether and how it is possible to make algorithm MC-UCB anytime.
Although we will not provide formal proofs of this result in this paper, we believe that setting a δ that evolves with the current time, as $\delta_t = t^{-7/2}$, is sufficient to make all the regret bounds of this paper hold with slightly modified constants. Some ideas on how to prove this result can be found in the article (Grover, 2009), and also (Auer et al., 2002) for something more specific to UCB algorithms.

7. Numerical experiment: Pricing of an Asian option

We consider the pricing problem of an Asian option introduced in (Glasserman et al., 1999) and later considered in (Kawai, 2010; Etoré and Jourdain, 2010). This uses a Black-Scholes model with strike C and maturity T. Let $(W(t))_{0\leq t\leq 1}$ be a Brownian motion that is discretized at d equidistant times $\{i/d\}_{1\leq i\leq d}$, which defines the vector $W \in \mathbb{R}^d$ with

components $W_i = W(i/d)$. The discounted payoff of the Asian option is defined as a function of W, by:

$$F(W) = \exp(-rT) \max\left[\frac{1}{d}\sum_{i=1}^{d} S_0 \exp\left[(r - \frac{1}{2}s_0^2)\frac{iT}{d} + s_0\sqrt{T}W_i\right] - C, 0\right], \qquad (9)$$

where S_0, r, and s_0 are constants, and the price is defined by the expectation $p = \mathbb{E}_W F(W)$.

We want to estimate the price p by Monte-Carlo simulations (by sampling on $W = (W_i)_{1\leq i\leq d}$). In order to reduce the variance of the estimated price, we can stratify the space of W. Glasserman et al. (1999) suggest to stratify according to a one dimensional projection of W, i.e., by choosing a projection vector $u \in \mathbb{R}^d$ and define the strata as the set of W such that $u \cdot W$ lies in intervals of $\mathbb{R}$. They further argue that the best direction for stratification is to choose $u = (0, \cdots, 0, 1)$, i.e., to stratify according to the last component W_d of W. Thus we sample W_d and then conditionally sample $W_1, ..., W_{d-1}$ according to a Brownian Bridge as explained in (Kawai, 2010). Note that this choice of stratification is also intuitive since W_d has the biggest exponent in the payoff (9), and thus the highest volatility. Kawai (2010) and Etoré and Jourdain (2010) also use the same direction of stratification.

Like in (Kawai, 2010) we consider 5 strata of equal weight. Since W_d follows a $\mathcal{N}(0, 1)$, the strata correspond to the 20-percentile of a normal distribution. The left plot of Figure 2 represents the cumulative distribution function of W_d and shows the strata in terms of percentiles of W_d. The right plot represents, in dot line, the curve $\mathbb{E}[F(W)|W_d = x]$ versus $\mathbb{P}(W_d < x)$ parameterized by x, and the box plot represents the expectation and standard deviations of $F(W)$ conditioned on each stratum. We observe that this stratification produces an important heterogeneity of the standard deviations per stratum, which indicates that a stratified sampling would be profitable compared to a crude Monte-Carlo sampling.

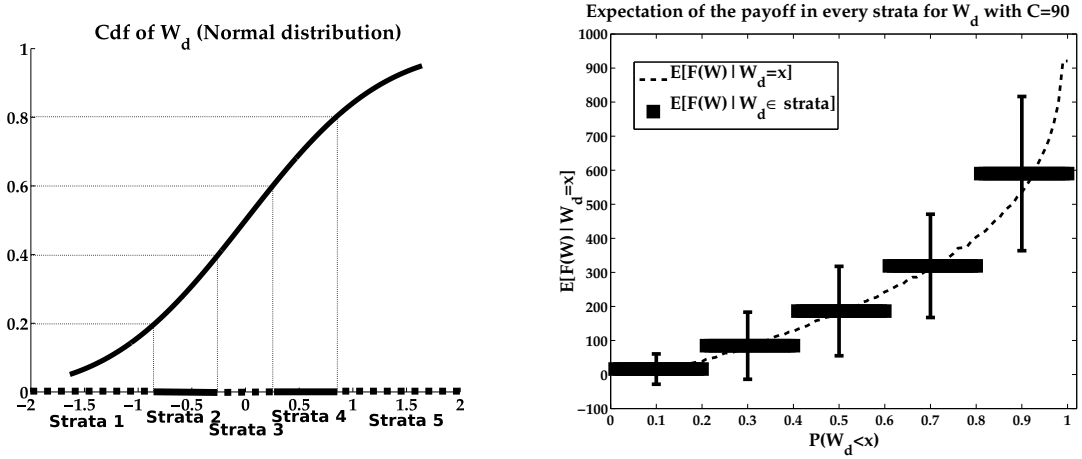

Figure 2: Left: Cdf of W_d and the definition of the strata. Right: expectation and standard deviation of $F(W)$ conditioned on each stratum for a strike $C = 90$.

We choose the same numerical values as Kawai (2010): $S_0 = 100$, $r = 0.05$, $s_0 = 0.30$, $T = 1$ and $d = 16$. Note that the strike C of the option has a direct impact on the variability of the strata. Indeed, the larger C, the more probable $F(W) = 0$ for strata with small W_d, and thus, the smaller $\lambda_{\min}$.

Our two main competitors are the SSAA algorithm of Etoré and Jourdain (2010) and GAFS-WL of Grover (2009). We did not compare to (Kawai, 2010) which aims at minimizing the computational time and not the loss considered here[12]. SSAA works in K_r rounds of length N_k where, at each round, it allocates proportionally to the empirical standard deviations computed in the previous rounds. Etoré and Jourdain (2010) report the asymptotic consistency of the algorithm whenever $\frac{k}{N_k}$ goes to 0 when k goes to infinity. Since their goal is not to obtain a finite-time performance, they do not mention how to calibrate the length and number of rounds in practice. We choose the same parameters as in their numerical experiments (Section 3.2.2 of (Etoré and Jourdain, 2010)) using 3 rounds. In this setting where we know the budget n at the beginning of the algorithm, GAFS-WL pulls each arm $a\sqrt{n}$ times and then pulls at time $t+1$ the arm k_{t+1} that maximizes $\frac{w_k \hat{\sigma}_{k,t}}{T_{k,t}}$. We set $a = 1$.

As mentioned in Subsection 6.4, an advantage of our algorithm is that it requires a single parameter to tune. We chose $b = 1000 \log(n)$ where 1000 is a high-probability range of the variables (see right plot of Figure 2). Table 7 reports the performance of MC-UCB, GAFS-WL, SSAA, and the uniform strategy, for different values of strike C i.e., for different values of $\lambda_{\min}^{-1}$ and $\Sigma_{w,2}/\Sigma_w^2 = \frac{\sum w_k \sigma_k^2}{(\sum_k w_k \sigma_k)^2}$. The total budget is $n = 10^5$. The results are averaged on 50000 trials. We notice that MC-UCB outperforms the uniform strategy, SSAA, and GAFS-WL. Note however that, in the case of GAFS-WL strategy, the small gain could come from the fact that there are more parameters in MC-UCB, and that we were thus able to adjust them (even if we kept the same parameters for the three values of C). Note however that for small (but non-zero) values of $\lambda_{\min}$, we proved in Appendix E that algorithm GAFS-WL was arbitrarily inefficient.

C	$\frac{1}{\lambda_{\min}}$	$\Sigma_{w,2}/\Sigma_w^2$	Uniform	SSAA	GAFS-WL	MC-UCB
60	6.18	1.06	$2.52\ 10^{-2}$	$5.87\ 10^{-3}$	$8.25\ 10^{-4}$	$7.29\ 10^{-4}$
90	15.29	1.24	$3.32\ 10^{-2}$	$6.14\ 10^{-3}$	$8.58\ 10^{-4}$	$8.07\ 10^{-4}$
120	744.25	3.07	$3.56\ 10^{-2}$	$6.22\ 10^{-3}$	$9.89\ 10^{-4}$	$9.28\ 10^{-4}$

Table 1: Characteristics of the distributions ($\lambda_{\min}^{-1}$ and $\Sigma_{w,2}/\Sigma_w^2$) and regret of the Uniform, SSAA, and MC-UCB strategies, for different values of the strike C.

In the left plot of Figure 3, we plot the rescaled true regret $\bar{R}_n n^{3/2}$, averaged over 50000 trials, as a function of n, where n ranges from 50 to 5000. The value of the strike is $C = 120$. Again, we notice that MC-UCB performs better than Uniform and SSAA because it adapts faster to the distributions of the strata. But it performs very similarly to GAFS-WL. In addition, it seems that the true regret of Uniform and SSAA grows faster than the rate $n^{3/2}$, whereas MC-UCB, as well as GAFS-WL, grow with this rate. The right plot focuses on the MC-UCB algorithm and rescales the $y-$axis to observe the variations of its rescaled true regret more accurately. The curve grows first and then stabilizes. This could correspond to the two regimes discussed previously.

12. In that article, the computational costs for each stratum vary, i.e. it is faster to sample in some strata than in others, and the aim of the article is to minimize the global computational cost while achieving a given performance.

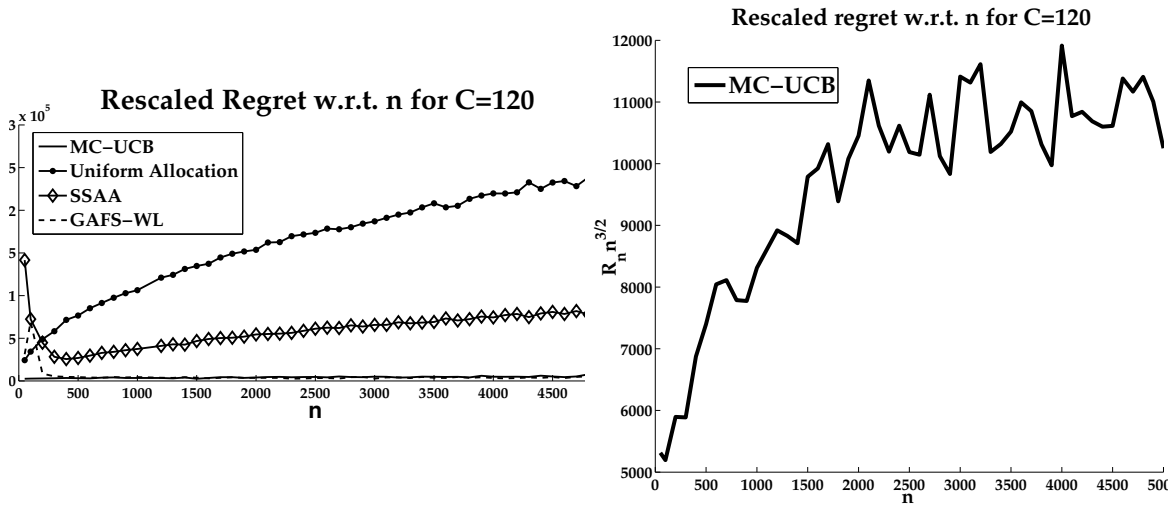

Figure 3: Left: Rescaled true regret ($\bar{R}_n n^{3/2}$) of the Uniform, SSAA, and MC-UCB strategies. Right: zoom on the rescaled regret for MC-UCB that illustrates the two regimes.

8. Conclusions

We provide a finite-time analysis for stratified sampling for Monte-Carlo in the case of fixed strata. We reported two bound on the expectation of the pseudo-regret: (i) a distribution dependent bound of order $\widetilde{O}(n^{-3/2}\lambda_{\min}^{-5/2})$ which is of interest when n is large compared to a measure of disparity $\lambda_{\min}^{-1}$ of the standard deviations (*stationary regime*), and (ii) a distribution free bound of order $\widetilde{O}(n^{-4/3})$ which is of interest when n is small compared to $\lambda_{\min}^{-1}$ (*transitory regime*). We also link the expectation of the pseudo-loss to the mean-squared error of algorithm MC-UCB and provide also problem dependent and problem independent bounds. An immediate consequence is the asymptotic convergence of the variance of our estimate to the optimal variance that requires the knowledge of the standard deviations per stratum.

We also provide the first problem independent (minimax) lower bound on the expectation of the pseudo-regret for this problem. Interestingly, the problem independent bound on expectation of the pseudo-regret of MC-UCB matches this lower-bound, both in terms of number of strata K and in terms of budget n. This means that algorithm MC-UCB is minimax-optimal in terms of pseudo-regret.

Possible directions for future work include: (i) making the MC-UCB algorithm anytime (i.e. not requiring the knowledge of n) and (ii) deriving distribution-dependent lower-bound for this problem and (iii) proposing efficient ways to stratify the space depending on the regularity of the function.

Supplementary material for the paper : Finite Time Analysis of Stratified Sampling for Monte Carlo

Appendix A. Proof of Theorem 1

Let us write the proof of the lower bound using the terminology of multi-armed bandits. Each arm k represents a stratum and the distribution associated to this arm is defined as the distribution of the noisy samples of the function collected when sampling uniformly on the strata.

Let us choose $\mu < 1/2$ and $\alpha = \frac{\mu}{2}$. Consider $2K$ Bernoulli bandits (i.e., $2K$ strata where the samples follow Bernoulli distributions) where the K first bandits have parameter $(\mu_k)_{1\leq k\leq K}$ and the K last ones have parameter $1/2$. The μ_k take values in $\{\mu-\alpha, \mu, \mu+\alpha\}$.

Define $\sigma^2 = \mu(1-\mu)$ the variance of a Bernoulli of parameter μ, and is such that $\sqrt{\frac{1}{2}\mu} \leq \sigma \leq \sqrt{\mu}$. We wite $\sigma_{-\alpha}$ and $\sigma_{+\alpha}$ the two other standard deviations, and notice that $\frac{1}{2}\sqrt{\mu} \leq \sigma_{-\alpha} \leq \sqrt{\mu}$, and $\sqrt{\frac{1}{2}\mu} \leq \sigma_{+\alpha} \leq \sqrt{\mu}$.

We consider the 2^K bandit environments $M(\epsilon)$ (characterized by $\epsilon = (\epsilon_k)_{1\leq k\leq K} \in \{-1,+1\}^K$) defined by $(\mu_k = \mu + \epsilon_k\alpha)_{1\leq k\leq K}$. We write $\mathbb{P}_\epsilon$ the probability with respect to the environment $M(\epsilon)$ at time n. We also write $M(\sigma)$ the environment defined by all K first arms having a parameter σ, and write $\mathbb{P}_\sigma$ the associated probability at time n.

The optimal oracle allocation for environment $M(\epsilon)$ is to play arm $k \leq K$, $t_k(\epsilon) = \frac{\sigma_{\epsilon_k\alpha}}{\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2}n$ times and arm $k > K$, $t_k(\epsilon) = \frac{1/2}{\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2}n$ times. The corresponding quadratic error of the resulting estimate is $l(\epsilon) = \frac{(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2)^2}{(2K)^2 n}$. For the environment $M(\sigma)$, the optimal oracle allocation is to play arm $k \leq K$, $t(\sigma) = \frac{\sigma}{K\sigma+K/2}n$ times (and arm $k > K$, $t_2(\sigma) = \frac{1/2}{K\sigma+K/2}n$ times).

Consider deterministic algorithms first (extension to randomized algorithms will be discussed later). An algorithm is a set (for all $t = 1$ to $n-1$) of mappings from any sequence $(r_1, \ldots, r_t) \in \{0,1\}$ of t observed samples (where $r_s \in \{0,1\}$ is the sample observed at the s-th round) to the choice of an arm $I_{t+1} \in \{1, \ldots, 2K\}$. Write $T_k(r_1, \ldots, r_n)$ the (random variable) corresponding to the number of pulls of arm k up to time n. We thus have $n = \sum_{k=1}^{2K} T_k$.

Now, consider the set of algorithms that know that the K first arms have parameter $\mu_k \in \{\mu-\alpha, \mu, \mu+\alpha\}$, and that also know that the K last arms have their parameters in $\{1/4, 3/4\}$. Given this knowledge, an optimal algorithm will not pull any arm $k \leq K$ more than $\left(\frac{\sigma_{+\alpha}}{K\sigma_{-\alpha}+\sqrt{3}K/4}\right)n$ times. Indeed, the optimal oracle allocation in *all* such environments allocates less than $\left(\frac{\sigma_{+\alpha}}{K\sigma_{-\alpha}+\sqrt{3}K/4}\right)n$ samples to each arm $k \leq K$. In addition, since the samples of all arms are independent, a sample collected from arm k does not provide any information about the relative allocations among the other arms. Thus, once an arm has been pulled as many times as recommended by the optimal oracle strategy, there is no

need to allocate more samples to that arm. Writing $\mathbb{A}$ the class of all algorithms that do not know the set of possible environments, $\mathbb{A}_\epsilon$ the class of algorithms that know the set of possible environments $M(\epsilon)$ and $\mathbb{A}_{opt}$ the subclass of $\mathbb{A}_\epsilon$ that pull all arms $k \leq K$ less than $\left(\frac{\sigma_{+\alpha}}{K\sigma_{-\alpha}+\sqrt{3}K/4}\right)n$ times, we have

$$\inf_{\mathbb{A}} \sup_{M(\epsilon)} \mathbb{E}R_n \geq \inf_{\mathbb{A}_\epsilon} \sup_{M(\epsilon)} \mathbb{E}R_n = \inf_{\mathbb{A}_{opt}} \sup_{M(\epsilon)} \mathbb{E}R_n,$$

where the first inequality comes from the fact that algorithms in $\mathbb{A}_\epsilon$ possess more information than those in $\mathbb{A}$, which they can use or not. Thus $\mathbb{A} \subset \mathbb{A}_\epsilon$.

Now for any $\epsilon = (\epsilon_1, \ldots, \epsilon_K)$, define the events

$$\Omega_\epsilon = \{\omega : \forall \mathcal{U} \subset \{1, \ldots, K\} : |\mathcal{U}| \leq \frac{K}{3} \text{ and } \forall k \in \mathcal{U}^c, \epsilon_k T_k \geq \epsilon_k t(\sigma)\}.$$

Note that by definition

$$\Omega_\epsilon = \bigcup_{p=1}^{\frac{K}{3}} \bigcup_{\mathcal{U}\subset\{1,\ldots,K\}:|\mathcal{U}|=p} \left\{ \Big\{ \bigcap_{k\in\mathcal{U}} \{\epsilon_k T_k < \epsilon_k t(\sigma)\} \Big\} \bigcap \Big\{ \bigcap_{k\in\mathcal{U}^C} \{\epsilon_k T_k \geq \epsilon_k t(\sigma)\} \Big\} \right\}.$$

By the sub-additivity of the probabilities, we have

$$\mathbb{P}_\sigma(\Omega_\epsilon) \leq \sum_{p=1}^{\frac{K}{3}} \sum_{\mathcal{U}\subset\{1,\ldots,K\}:|\mathcal{U}|=p} \mathbb{P}\left[\left\{ \Big\{ \bigcap_{k\in\mathcal{U}} \{\epsilon_k T_k < \epsilon_k t(\sigma)\} \Big\} \bigcap \Big\{ \bigcap_{k\in\mathcal{U}^C} \{\epsilon_k T_k \geq \epsilon_k t(\sigma)\} \Big\} \right\} \right].$$

The events $\left\{ \Big\{ \bigcap_{k\in\mathcal{U}} \{\epsilon_k T_k < \epsilon_k t(\sigma)\} \Big\} \bigcap \Big\{ \bigcap_{k\in\mathcal{U}^C} \{\epsilon_k T_k \geq \epsilon t(\sigma)\} \Big\} \right\}$ are disjoint for different ϵ, and form a partition of the space, thus $\sum_\epsilon \mathbb{P}_\sigma \left[\left\{ \Big\{ \bigcap_{k\in\mathcal{U}} \{\epsilon_k T_k < \epsilon_k t(\sigma)\} \Big\} \bigcap \Big\{ \bigcap_{k\in\mathcal{U}^C} \{\epsilon T_k \geq \epsilon_k t(\sigma)\} \Big\} \right\} \right] = 1$.

We deduce that

$$\begin{aligned}
\sum_\epsilon \mathbb{P}_\sigma(\Omega_\epsilon) &\leq \sum_\epsilon \sum_{p=1}^{\frac{K}{3}} \sum_{\mathcal{U}\subset\{1,\ldots,K\}:|\mathcal{U}|=p} \mathbb{P}_\sigma \left[\left\{ \Big\{ \bigcap_{k\in\mathcal{U}} \{\epsilon T_k < \epsilon_k t(\sigma)\} \Big\} \bigcap \Big\{ \bigcap_{k\in\mathcal{U}^C} \{\epsilon_k T_k \geq \epsilon_k t(\sigma)\} \Big\} \right\} \right] \\
&= \sum_{p=1}^{\frac{K}{3}} \sum_{\mathcal{U}\subset\{1,\ldots,K\}:|\mathcal{U}|=p} \sum_\epsilon \left[\left\{ \Big\{ \bigcap_{k\in\mathcal{U}} \{\epsilon_k T_k < \epsilon_k t(\sigma)\} \Big\} \bigcap \Big\{ \bigcap_{k\in\mathcal{U}^C} \{\epsilon T_k \geq \epsilon_k t(\sigma)\} \Big\} \right\} \right] \\
&= \sum_{p=1}^{\frac{K}{3}} \sum_{\mathcal{U}\subset\{1,\ldots,K\}:|\mathcal{U}|=p} 1 \\
&= \sum_{p=1}^{\frac{K}{3}} \left(\begin{array}{c} K \\ p \end{array} \right).
\end{aligned}$$

Since there are 2^K environments ϵ, we have

$$\min_{\epsilon} \mathbb{P}_\sigma(\Omega_\epsilon) \leq \frac{1}{2^K} \sum_{\epsilon} \mathbb{P}_\sigma(\Omega_\epsilon) \leq \frac{1}{2^K} \sum_{p=1}^{\frac{K}{3}} \left(\begin{array}{c} K \\ p \end{array} \right).$$

Note that $\frac{1}{2^K} \sum_{p=1}^{\frac{K}{3}} \left(\begin{array}{c} K \\ p \end{array} \right) = \mathbb{P}(\sum_{k=1}^K X_k \leq \frac{K}{3})$ where $(X_1, \ldots, X_K)$ are K independent Bernoulli random variables of parameter $1/2$. By Chernoff-Hoeffding's inequality, we have $\mathbb{P}(\sum_{k=1}^K X_k \leq \frac{K}{3}) = \mathbb{P}(\frac{1}{K}\sum_{k=1}^K X_k - \frac{1}{2} \leq \frac{K}{6}) \leq \exp(-K/72)$. Thus there exists $\epsilon_{\min}$ such that $\mathbb{P}_\sigma(\Omega_{\epsilon_{\min}}) \leq \exp(-K/72)$.

Let us write $p = \mathbb{P}_{\epsilon_{\min}}(\Omega_{\epsilon_{\min}})$ and $p_\sigma = \mathbb{P}_\sigma(\Omega_{\epsilon_{\min}})$. Let $kl(a,b) = a\log(\frac{a}{b}) + (1-a)\log(\frac{1-a}{1-b})$ denote the KL for Bernoulli distributions with parameters a and b. Note that because $\forall \Omega$, $KL(\mathbb{P}_{\epsilon_{\min}}(.|\Omega), \mathbb{P}_\sigma(.|\Omega)) \geq 0$, we have

$$kl(p, p_\sigma) \leq KL(\mathbb{P}_{\epsilon_{\min}}, \mathbb{P}_\sigma).$$

From that we deduce that $p(\log(p) - \log(p_\sigma)) + (1-p)(\log(1-p) - \log(1-p_\sigma)) \leq KL(\mathbb{P}_{\epsilon_{\min}}, \mathbb{P}_\sigma)$, which leads to

$$p \leq \max(\frac{36}{K}\Big(KL(\mathbb{P}_{\epsilon_{\min}}, \mathbb{P}_\sigma)\Big), \exp(-K/72)). \tag{10}$$

Let us now consider any environment (ϵ). Let $R_t = (r_1, \ldots, r_t)$ be the sequence of observations, and let $\mathbb{P}_\epsilon^t$ be the law of R_t for environment $M(\epsilon)$. Note first that $\mathbb{P}_\epsilon = \mathbb{P}_\epsilon^n$. Adapting the chain rule for Kullback-Leibler divergence, we get

$$\begin{aligned}
&KL(\mathbb{P}_\epsilon^n, \mathbb{P}_\sigma^n) \\
&= KL(\mathbb{P}_\epsilon^1, \mathbb{P}_\sigma^1) + \sum_{t=2}^n \sum_{R_{t-1}} \mathbb{P}_\epsilon^{t-1}(R_{t-1}) KL(\mathbb{P}_\epsilon^t(.|R_{t-1}), \mathbb{P}_\sigma^t(.|R_t)) \\
&= KL(\mathbb{P}_\sigma^1, \mathbb{P}_\epsilon^1) + \sum_{t=2}^n \Big[\sum_{R_{t-1}|\epsilon_{I_t}=+1} \mathbb{P}_\sigma^{t-1}(R_{t-1}) kl(\mu+\alpha, \mu) + \sum_{R_{t-1}|\epsilon_{I_t}=-1} \mathbb{P}_\sigma^{t-1}(R_{t-1}) kl(\mu-\alpha, \mu) \Big] \\
&= kl(\mu-\alpha, \mu)\mathbb{E}_\epsilon[\sum_{k:\epsilon_k=-1} T_k] + kl(\mu+\alpha, \mu)\mathbb{E}_\epsilon[\sum_{k:\epsilon_k=+1} T_k].
\end{aligned}$$

We thus have, using the property that $kl(a,b) \leq \frac{(a-b)^2}{b(1-b)}$,

$$\begin{aligned}
KL(\mathbb{P}_\epsilon, \mathbb{P}_\sigma) &= kl(\mu-\alpha, \mu)\mathbb{E}_\epsilon[\sum_{k:\epsilon_k=-1} T_k] + kl(\mu+\alpha, \mu)\mathbb{E}_\epsilon[\sum_{k:\epsilon_k=+1} T_k] \\
&\leq \mathbb{E}_\sigma[\sum_{k\leq K} T_k] \frac{\alpha^2}{\mu(1-\mu)} \\
&= E_\sigma[\sum_{k\leq K} T_k] \frac{\alpha^2}{\sigma^2}.
\end{aligned}$$

Note that for an algorithm in $\mathbb{A}_{opt}$, we have $\sum_{k=1}^{K} T_k \leq T_k \leq K\Big(\frac{\sigma_{+\alpha}}{K\sigma_{-\alpha}+\sqrt{3}K/4}\Big)n$. Since $\alpha = \frac{\mu}{2}$ and $0 < \mu \leq \frac{1}{2}$ we have

$$\begin{aligned} KL(\mathbb{P}_\epsilon, \mathbb{P}_\sigma) &\leq \Big(K\frac{\sigma_{+\alpha}}{K\sigma_{-\alpha}+\sqrt{3}K/4}\Big)\frac{\alpha^2}{\sigma^2}n \\ &\leq 4\sigma_{+\alpha}\frac{\alpha^2}{\sigma^2}n \\ &\leq 8\frac{\alpha^2}{\sigma}n, \end{aligned}$$

We thus deduce using Equation 10

$$\begin{aligned} \mathbb{P}_{\epsilon_{\min}}(\Omega_{\epsilon_{\min}}) = p &\leq \max(\frac{18}{K}\Big(KL(\mathbb{P}_{\epsilon_{\min}}, \mathbb{P}_\sigma)\Big), \exp(-K/72)) \\ &\leq \frac{144}{K}\frac{\alpha^2}{\sigma}n. \end{aligned}$$

Now choose $\sigma \leq \frac{1}{7}(\frac{K}{n})^{1/3}$ (as $\alpha = \frac{\mu}{2} = \frac{\sigma^2}{2}$). Note that this implies that $\mathbb{P}_{\epsilon_{\min}}(\Omega_{\epsilon_{\min}}) \leq \frac{1}{2}$.

Let $\omega \in \Omega^c_{\epsilon_{\min}}$. We know that for ω, there are at least $\frac{K}{3}$ arms among the K first which are not pulled correctly: either $\frac{K}{6}$ arms among the arms with parameter $\mu - \alpha$ or among the arms with parameter $\mu + \alpha$ are not pulled correctly. Assume that for this fixed ω, there are $\frac{K}{6}$ arms among the arms with parameter $\mu - \alpha$ which are not pulled correctly. Let $\mathcal{U}(\omega)$ be this subset of arms.

We write $\Delta T = \sum_{k\in\mathcal{U}} T_k - \frac{K}{6}t(\sigma_{-\alpha})$ the number of times those arms are over pulled. Note that on ω we have $\Delta T \geq \frac{K}{6}t(\sigma) - t(\sigma_{-\alpha})$. We have

$$\begin{aligned} \Delta T = \frac{K}{6}t(\sigma) - \frac{K}{6}t(\sigma_{-\alpha}) &= \frac{1}{6}\frac{K\sigma}{K\sigma + K/2}n - \frac{1}{6}\frac{K\sigma_{-\alpha}}{\sum_{i=1}^{K}\sigma_{\epsilon_i\alpha} + K/2}n \\ &\geq \frac{1}{6}\frac{K\sigma}{K\sigma + K/2}n - \frac{1}{6}\frac{K\sigma/\sqrt{2}}{\sqrt{3}K\sigma/\sqrt{2} + K/2}n \\ &\geq \frac{1}{6}\frac{1}{K\sigma + K/2}\frac{1}{\sqrt{3}K\sigma/\sqrt{2} + K/2}\Big(K^2\sigma/2 - K^2\sigma/2\sqrt{2}\Big)n \\ &\geq \frac{1}{2}(1 - 1/\sqrt{2})\sigma n \\ &\geq \frac{1}{35}K^{1/3}n^{2/3} \end{aligned}$$

Thus on ω, the regret is such that

$$\begin{aligned}
R_{n,\epsilon_{\min}}(\omega) &\geq \sum_{k=1}^{3K} \frac{w_k^2\sigma_k^2}{T_k(\omega)} - \frac{1}{(2K)^2}\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha} + K/2\right)^2}{n} \\
&\geq \sum_{k\in\mathcal{U}(\omega)} \frac{w_k^2\sigma_k^2}{T_k(\omega)} + \sum_{k\in\mathcal{U}(\omega)^C} \frac{w_k^2\sigma_k^2}{T_k(\omega)} - \frac{1}{(2K)^2}\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha} + K/2\right)^2}{n} \\
&\geq \frac{1}{K^2}\frac{K}{6}\frac{\sigma_{-\alpha}^2}{t_k(\sigma_{-\alpha}) + 6\Delta T/K} + \frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha} - K\sigma_{-\alpha}/6 + K/2\right)^2}{(2K - K/6)^2(n-\Delta T)} - \frac{1}{(2K)^2}\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha} + K/2\right)^2}{n} \\
&\geq \frac{1}{(2K)^2}\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha} + K/2\right)^2}{n}\frac{1 + \left(\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)\Delta T}{\left(K\sigma_{-\alpha}/6\right)n} - \frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)\Delta T}{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}-K\sigma_{-\alpha}/6+K/2\right)n}\right)}{\left(1 + \frac{6\Delta T\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)}{K\sigma_{-\alpha}n}\right)\left(1 - \frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)\Delta T}{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}-K\sigma_{-\alpha}/6+K/2\right)n}\right)} \\
&\quad - \frac{1}{(2K)^2}\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha} + K/2\right)^2}{n} \\
&\geq \frac{1}{(2K)^2}\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha} + K/2\right)^2}{n}\frac{\left(\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)\Delta T}{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}-K\sigma_{-\alpha}/6+K/2\right)n}\right)\left(\frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)\Delta T}{\left(K\sigma_{-\alpha}/6\right)n}\right)}{\left(1 + \frac{6\Delta T\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)}{K\sigma_{-\alpha}n}\right)\left(1 - \frac{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}+K/2\right)\Delta T}{\left(\sum_{i=1}^K \sigma_{\epsilon_i\alpha}-K\sigma_{-\alpha}/6+K/2\right)n}\right)} \\
&\geq C\frac{(\Delta T)^2}{n^3\sigma} \\
&\geq C\frac{K^{1/3}}{n^{4/3}},
\end{aligned}$$

where C is a numerical constant. Note that for events ω where there are $\frac{K}{6}$ arms among the arms with parameter $\mu+\alpha$ which are not pulled correctly, the same result holds.

Note finally that $\mathbb{P}(\Omega^c_{\epsilon_{\min}}) \geq 1/2$. We thus have that the regret is bigger than

$$\begin{aligned}
\mathbb{E}R_{n,\epsilon_{\min}} &\geq \sum_{\omega\in\Omega^c_{\epsilon_{\min}}} R_{n,\epsilon_{\min}}(\omega)\mathbb{P}_{\epsilon_{\min}}(\omega) \\
&\geq \sum_{\omega\in\Omega^c_{\epsilon_{\min}}} C\frac{K^{1/3}}{n^{4/3}}\mathbb{P}_{\epsilon_{\min}}(\omega) \\
&\geq \frac{1}{2}C\frac{K^{1/3}}{n^{4/3}},
\end{aligned}$$

which proves the lower bound for deterministic algorithms. Now the extension to randomized algorithms is straightforward: any randomized algorithm can be seen as a static (i.e., does not depend on samples) mixture of deterministic algorithms (which can be defined before the game starts). Each deterministic algorithm satisfies the lower bound above in expectation, thus any static mixture does so too.

Appendix B. Main technical tools for the regret and pseudo-regret bounds

B.1 The main tool: a high probability bound on the standard deviations

Upper bound on the standard deviation: The upper confidence bounds $B_{k,t}$ used in the MC-UCB algorithm is motivated by Theorem 10 in (Maurer and Pontil, 2009) (a variant of this result is also reported in (Audibert et al., 2009)). We extend this result to sub-Gaussian random variables.

Lemma 11 *Let Assumption 1 hold and $n \geq 2$. Define the following event*

$$\xi = \xi_{K,n}(\delta) = \bigcap_{1\leq k\leq K,\, 2\leq t\leq n} \left\{ \left| \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\Big(X_{k,i} - \frac{1}{t}\sum_{j=1}^{t} X_{k,j}\Big)^2} - \sigma_k \right| \leq 2a\sqrt{\frac{\log(2/\delta)}{t}} \right\}, \tag{11}$$

where $a = \sqrt{2c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}} n^{1/2}$. *Then* $\Pr(\xi) \geq 1 - 2nK\delta$.

Note that the first term in the absolute value in Equation 11 is the empirical standard deviation of arm k computed as in Equation 7 for t samples. The event ξ plays an important role in the proofs of this section and a number of statements will be proved on this event.

Proof

Step 1. Truncating sub-Gaussian variables. We want to characterize the mean and variance of the variables $X_{k,t}$ given that $|X_{k,t} - \mu_k| \leq \sqrt{c_1\log(c_2/\delta)}$. For any positive random variable Y and any $b \geq 0$, $\mathbb{E}(Y\mathbb{I}\{Y > b\}) = \int_b^\infty \mathbb{P}(Y > \epsilon)d\epsilon + b\mathbb{P}(Y > b)$. If we take $b = c_1\log(c_2/\delta)$ and use Assumption 1, we obtain:

$$\begin{aligned}
\mathbb{E}\Big[|X_{k,t} - \mu_k|^2\mathbb{I}\left\{|X_{k,t} - \mu_k|^2 > b\right\}\Big] &= \int_b^{+\infty} \mathbb{P}\big(|X_{k,t} - \mu_k|^2 > \epsilon\big)d\epsilon + b\mathbb{P}(|X_{k,t} - \mu_k|^2 > b) \\
&\leq \int_b^{+\infty} c_2\exp(-\epsilon/c_1)d\epsilon + bc_2\exp(-b/c_1) \\
&\leq c_1\delta + c_1\log(c_2/\delta)\delta \\
&\leq c_1\delta(1 + \log(c_2/\delta)).
\end{aligned}$$

We have $\mathbb{E}\Big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{|X_{k,t} - \mu_k|^2 > b\right\}\Big] + \mathbb{E}\Big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{|X_{k,t} - \mu_k|^2 \leq b\right\}\Big] = \sigma_k^2$, which, combined with the previous equation, implies that

$$\begin{aligned}
\Big|\mathbb{E}\Big[|X_{k,t} - \mu_k|^2 \;\big|\; |X_{k,t} - \mu_k|^2 \leq b\Big] - \sigma_k^2\Big| &= \frac{\Big|\mathbb{E}\Big[\Big((X_{k,t} - \mu_k)^2 - \sigma_k^2\Big)\mathbb{I}\left\{|X_{k,t} - \mu_k|^2 > b\right\}\Big]\Big|}{\mathbb{P}\Big(|X_{k,t} - \mu_k|^2 \leq b\Big)} \\
&\leq \frac{c_1\delta(1 + \log(c_2/\delta)) + \delta\sigma_k^2}{1-\delta}.
\end{aligned} \tag{12}$$

Note also that Cauchy-Schwartz inequality implies

$$\left|\mathbb{E}\Big[\Big(X_{k,t}-\mu_k\Big)\mathbb{I}\left\{|X_{k,t}-\mu_k|^2>b\right\}\Big]\right| \leq \sqrt{\mathbb{E}\Big[(X_{k,t}-\mu_k)^2\mathbb{I}\left\{|X_{k,t}-\mu_k|^2>b\right\}\Big]} \leq \sqrt{c_1\delta(1+\log(c_2/\delta))}.$$

Now, notice that $\mathbb{E}\Big[X_{k,t}\mathbb{I}\left\{|X_{k,t}-\mu_k|^2>b\right\}\Big]+\mathbb{E}\Big[X_{k,t}\mathbb{I}\left\{|X_{k,t}-\mu_k|^2\leq b\right\}\Big]=\mu_k$, which, combined with the previous result and using $n\geq K\geq 2$, implies that

$$|\tilde{\mu}_k-\mu_k| = \frac{\left|\mathbb{E}\Big[\Big(X_{k,t}-\mu_k\Big)\mathbb{I}\left\{|X_{k,t}-\mu_k|^2>b\right\}\Big]\right|}{\mathbb{P}\Big(|X_{k,t}-\mu_k|^2\leq b\Big)} \leq \frac{\sqrt{c_1\delta(1+\log(c_2/\delta))}}{1-\delta}, \tag{13}$$

where $\tilde{\mu}_k \overset{\text{def}}{=} \mathbb{E}\Big[X_{k,t} \mid |X_{k,t}-\mu_k|^2\leq b\Big] = \frac{\mathbb{E}\Big[X_{k,t}\mathbb{I}\left\{|X_{k,t}-\mu_k|^2\leq b\right\}\Big]}{\mathbb{P}\Big(|X_{k,t}-\mu_k|^2\leq b\Big)}$.

We note $\tilde{\sigma}_k^2 \overset{\text{def}}{=} \mathbb{V}\Big[X_{k,t} \mid |X_{k,t}-\mu_k|^2\leq b\Big] = \mathbb{E}\Big[|X_{k,t}-\mu_k|^2 \mid |X_{k,t}-\mu_k|^2\leq b\Big]-(\mu_k-\tilde{\mu_k})^2$. From Equations 12 and 13, we derive

$$\begin{aligned}|\tilde{\sigma}_k^2-\sigma_k^2| &\leq \left|\mathbb{E}\Big[|X_{k,t}-\mu_k|^2 \mid |X_{k,t}-\mu_k|^2\leq b\Big]-\sigma_k^2\right|+|\tilde{\mu}_k-\mu_k|^2\\ &\leq \frac{c_1\delta(1+\log(c_2/\delta))+\delta\sigma_k^2}{1-\delta}+\frac{c_1\delta(1+\log(c_2/\delta))}{(1-\delta)^2}\\ &\leq \frac{2c_1\delta(1+\log(c_2/\delta))+\delta\sigma_k^2}{(1-\delta)^2},\end{aligned}$$

from which we deduce, because $\sigma_k^2\leq c_1c_2$

$$|\tilde{\sigma_k}-\sigma_k| \leq \frac{\sqrt{2c_1\delta(1+c_2+\log(c_2/\delta))}}{1-\delta}. \tag{14}$$

Step 2. Application of large deviation inequalities.

Let $\xi_1=\xi_{1,K,n}(\delta)$ be the event:

$$\xi_1 = \bigcap_{1\leq k\leq K,\ 1\leq t\leq n}\left\{|X_{k,t}-\mu_k|\leq\sqrt{c_1\log(c_2/\delta)}\right\}.$$

Under Assumption 1, using a union bound, we have that the probability of this event is at least $1-nK\delta$.

We now recall Theorem 10 of (Maurer and Pontil, 2009):

Theorem 1 (Maurer and Pontil (2009)) *Let $(X_1,...,X_t)$ be $t\geq 2$ i.i.d. random variables of variance σ^2 and mean μ and such that $\forall i\leq t, X_i\in[a,a+c]$. Then with probability at least $1-\delta$:*

$$\left|\sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\Big(X_i-\frac{1}{t}\sum_{j=1}^{t}X_j\Big)^2}-\sigma\right| \leq 2c\sqrt{\frac{\log(2/\delta)}{t-1}}.$$

On ξ_1, the $\{X_{k,i}\}_i$, $1 \leq k \leq K$, $1 \leq i \leq t$ are t i.i.d. bounded random variables with standard deviation $\tilde{\sigma_k}$.

Let $\xi_2 = \xi_{2,K,n}(\delta)$ be the event:

$$\xi_2 = \bigcap_{1\leq k\leq K,\, 1\leq t\leq n} \left\{ \left| \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\Big(X_{k,i} - \frac{1}{t}\sum_{j=1}^{t} X_{k,j}\Big)^2} - \tilde{\sigma_k} \right| \leq 2\sqrt{c_1 \log(c_2/\delta)}\sqrt{\frac{\log(2/\delta)}{t-1}} \right\}.$$

Using Theorem 10 of (Maurer and Pontil, 2009) and a union bound, we deduce that $\Pr(\xi_1 \cap \xi_2) \geq 1 - 2nK\delta$.

Now, from Equation 14, we have on $\xi_1 \cap \xi_2$, for all $1 \leq k \leq K$, $2 \leq t \leq n$:

$$\begin{aligned}
\left| \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\Big(X_{k,i} - \frac{1}{t}\sum_{j=1}^{t} X_{k,j}\Big)^2} - \sigma_k \right| &\leq 2\sqrt{c_1 \log(c_2/\delta)}\sqrt{\frac{\log(2/\delta)}{t-1}} \\
&\quad + \frac{\sqrt{2c_1\delta(1 + c_2 + \log(c_2/\delta))}}{1-\delta} \\
&\leq 2\sqrt{2c_1 \log(c_2/\delta)}\sqrt{\frac{\log(2/\delta)}{t}} \\
&\quad + \frac{\sqrt{2c_1\delta(1 + c_2 + \log(c_2/\delta))}}{1-\delta},
\end{aligned}$$

from which we deduce Lemma 11 (since $\xi_1 \cap \xi_2 \subseteq \xi$ and $2 \leq t \leq n$). ■

We deduce the following corollary when the number of samples $T_{k,t}$ are random.

Corollary 12 *For any $k = 1, \ldots, K$ and $t = 2K, \ldots, n$, let $\{X_{k,i}\}_i$ be n i.i.d. random variables drawn from ν_k, satisfying Assumption 1. Let $T_{k,t}$ be any random variable taking values in $\{2, \ldots, n\}$. Let $\hat{\sigma}^2_{k,t}$ be the empirical variance computed from Equation 7. Then, on the event ξ, we have:*

$$|\hat{\sigma}_{k,t} - \sigma_k| \leq 2a\sqrt{\frac{\log(2/\delta)}{T_{k,t}}}\,. \tag{15}$$

B.2 Other important properties

A stopping time problem: We now draw a connection between the adaptive sampling and stopping time problems. We report the following proposition which is a type of Wald's Theorem for variance (see e.g. Resnick (1999)).

Proposition 13 *Let $\{\mathcal{F}_t\}$ be a filtration and X_t a $\mathcal{F}_t$-adapted sequence of i.i.d. random variables with variance σ^2. Assume that $\mathcal{F}_t$ and the σ-algebra generated by $\{X_i : i \geq t+1\}$ are independent and T is a stopping time w.r.t. $\mathcal{F}_t$ with a finite expected value. If $\mathbb{E}[X_1^2] < \infty$ then*

$$\mathbb{E}\left[\Big(\sum_{i=1}^{T} X_i - T\,\mu\Big)^2\right] = \mathbb{E}[T]\,\sigma^2. \tag{16}$$

Bound on $\mathbb{E}\big[|\hat{\mu}_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big]$. The next lemma provides a bound for the loss whenever the event ξ does not hold.

Lemma 14 *Let Assumption 1 holds. Then for every arm k:*

$$\mathbb{E}\big[|\hat{\mu}_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq 2c_1 n^2 K\delta(1+\log(c_2/2nK\delta)) \,.$$

Proof Since the arms have sub-Gaussian distribution, for any $1 \leq k \leq K$ and $1 \leq t \leq n$, we have

$$\mathbb{P}\big(|X_{k,t}-\mu_k|^2 \geq \epsilon\big) \leq c_2 \exp(-\epsilon/c_1) \,,$$

and thus by setting $\epsilon = c_1 \log(c_2/2nK\delta)$[13], we obtain

$$\mathbb{P}\big(|X_{k,t}-\mu_k|^2 \geq c_1 \log(c_2/2nK\delta)\big) \leq 2nK\delta \,.$$

We thus know that

$$\begin{aligned}
&\max_{\Omega/\mathbb{P}(\Omega)=2nK\delta} \mathbb{E}\big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{\Omega\right\}\big] \\
&\leq \int_{c_1\log(c_2/2nK\delta)}^{\infty} c_2\exp(-\epsilon/c_1)d\epsilon + c_1\log(c_2/2nK\delta)\mathbb{P}\Big(\Omega\Big) \\
&= 2c_1 nK\delta(1+\log(c_2/2nK\delta)) \,.
\end{aligned}$$

Since the event ξ^C has a probability at most $2nK\delta$, for any $1 \leq k \leq K$ and $1 \leq t \leq n$, we have

$$\mathbb{E}\big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq \max_{\Omega/\mathbb{P}(\Omega)=2nK\delta} \mathbb{E}\big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{\Omega\right\}\big] \leq 2c_1 nK\delta(1+\log(c_2/2nK\delta)) \,.$$

The claim follows from the fact that $\mathbb{E}\big[|\hat{\mu}_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq \sum_{t=1}^{n}\mathbb{E}\big[|X_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq 2c_1 n^2 K\delta(1+\log(c_2/2nK\delta))$. ∎

B.3 Technical inequalities

Upper and lower bound on a: If $\delta = n^{-7/2}$, with $n \geq 4K \geq 8$

$$\begin{aligned}
a &= \sqrt{2c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}} n^{1/2} \\
&\leq \sqrt{7c_1(c_2+1)\log(n)} + \frac{1}{n^{3/2}}\sqrt{c_1(2+c_2)} \\
&\leq 2\sqrt{2c_1(c_2+2)\log(n)}.
\end{aligned}$$

We also have by just keeping the first term and choosing c_2 such that $c_2 \geq e\delta = en^{-7/2}$

$$\begin{aligned}
a &= \sqrt{2c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}} n^{1/2} \\
&\geq \sqrt{2c_1} \geq \sqrt{c_1}.
\end{aligned}$$

13. Note that we need to choose c_2 such that $c_2 \geq 2nK\delta = 2Kn^{-5/2}$ if $\delta = n^{-7/2}$.

Lower bound on $c(\delta)$ when $\delta = n^{-7/2}$: Since the arms have sub-Gaussian distribution, for any $1 \leq k \leq K$ and $1 \leq t \leq n$, we have

$$\mathbb{P}\big(|X_{k,t} - \mu_k|^2 \geq \epsilon\big) \leq c_2 \exp(-\epsilon/c_1) \ ,$$

We then have

$$\mathbb{E}\big[|X_{k,t} - \mu_k|^2\big] \leq \int_0^\infty c_2 \exp(-\epsilon/c_1) d\epsilon = c_2 c_1$$

We then have $\Sigma_w \leq \sqrt{c_2 c_1}$.
If $\delta = n^{-7/2}$, we obtain by using the lower bound on a that

$$\begin{aligned} c(\delta = n^{-7/2}) &= \Big(\frac{2a\sqrt{\log(2/\delta)}}{\Sigma_w + 4a\sqrt{\log(2/\delta)}}\Big)^{2/3} \\ &= \Big(\frac{1}{2} - \frac{1}{2}\frac{\Sigma_w}{\Sigma_w + 4a\sqrt{\log(2/\delta)}}\Big)^{2/3} \\ &\geq \Big(\frac{1}{2} - \frac{1}{2}\frac{\Sigma_w}{\Sigma_w + 4\sqrt{c_1 \log(n)}}\Big)^{2/3} \\ &\geq \Big(\frac{1}{2}\Big)^{2/3}\Big(\frac{\sqrt{c_1}}{\Sigma_w + \sqrt{c_1}}\Big)^{2/3} \geq \Big(\frac{1}{2K}\Big)^{2/3}\Big(\frac{1}{\sqrt{c_2}+1}\Big)^{2/3}, \end{aligned}$$

by using $\Sigma_w \leq \sqrt{c_2 c_1}$ for the last step.

Upper bound on $\mathbb{E}\big[|\hat{\mu}_{k,n} - \mu_k|^2 \mathbb{I}\left\{\xi^C\right\}\big]$ when $\delta = n^{-7/2}$: We get from Lemma 14 when $\delta = n^{-7/2}$ and when choosing c_2 such that $c_2 \geq 2nK\delta = 2Kn^{-5/2}$

$$\begin{aligned} \mathbb{E}\big[|\hat{\mu}_{k,n} - \mu_k|^2 \mathbb{I}\left\{\xi^C\right\}\big] &\leq 2c_1 n^2 K\delta(1 + \log(c_2/2nK\delta)) \\ &\leq 2c_1 K\big(1 + \frac{5}{2}(c_2+1)\log(n)\big)n^{-3/2} \\ &\leq 6c_1 K(c_2+1)\log(n) n^{-3/2}. \end{aligned}$$

Appendix C. Proof of Theorems 2 and Proposition 4

In this section, we first provide the proof for an important Lemma on the number of pulls of the arms, and then use the result to prove Theorem 2 and Proposition 4.

C.1 Problem dependent bound on the number of pulls

Lemma 15 *Let Assumption 1 hold. Let $0 < \delta \leq 1$ be arbitrary and and $n \geq 4K$. The difference between the allocation $T_{p,n}$ implemented by the MC-UCB algorithm described in Figure 1 and the optimal allocation rule $T^*_{p,n}$ has the following upper and lower bounds, on ξ (and thus with probability at least $1 - 2nK\delta$), for any arm $1 \leq p \leq K$:*

$$-12a\lambda_p \frac{\sqrt{\log(2/\delta)}}{\Sigma_w \lambda_{\min}^{3/2}}\sqrt{n} - 4K\lambda_p \leq T_{p,n} - T^*_{p,n} \leq 12a\frac{\sqrt{\log(2/\delta)}}{\Sigma_w \lambda_{\min}^{3/2}}\sqrt{n} + 4K \ . \qquad (17)$$

where $a = \sqrt{2c_1 \log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}} n^{1/2}$.

In Equation 17, the difference $T_{p,n} - T^*_{p,n}$ is bounded with $\widetilde{O}(\sqrt{n})$. This is directly linked to the parametric rate of convergence of the estimation of σ_k, which is of order $1/\sqrt{n}$. Note that Equation 17 also shows the inverse dependency on the smallest proportion $\lambda_{\min}$.

Proof [Lemma 15] The proof consists of the following three main steps.

Step 1. Properties of the algorithm. Recall the definition of the upper bound used in MC-UCB when $t > 2K$:

$$B_{q,t+1} = \frac{w_q}{T_{q,t}} \left(\hat{\sigma}_{q,t} + 2a \sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right), \qquad 1 \leq q \leq K \ .$$

From Corollary 12, we obtain the following upper and lower bounds for $B_{q,t+1}$ on ξ:

$$\frac{w_q \sigma_q}{T_{q,t}} \leq B_{q,t+1} \leq \frac{w_q}{T_{q,t}} \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right). \tag{18}$$

Let $t+1 > 2K$ be the time at which a given arm k is pulled for the last time, i.e., $T_{k,t} = T_{k,n} - 1$ and $T_{k,(t+1)} = T_{k,n}$. Note that as $n \geq 4K$, there is at least one arm k such that this happens, i.e. such that it is pulled after the initialization phase. Since $\mathcal{A}_{MC-UCB}$ chooses to pull arm k at time $t+1$, we have for any arm p

$$B_{p,t+1} \leq B_{k,t+1} \ . \tag{19}$$

From Equation 18 and the fact that $T_{k,t} = T_{k,n} - 1$, we obtain

$$B_{k,t+1} \leq \frac{w_k}{T_{k,t}} \left(\sigma_k + 4a \sqrt{\frac{\log(2/\delta)}{T_{k,t}}} \right) = \frac{w_k}{T_{k,n}-1} \left(\sigma_k + 4a \sqrt{\frac{\log(2/\delta)}{T_{k,n}-1}} \right). \tag{20}$$

Using the lower bound in Equation 18 and the fact that $T_{p,t} \leq T_{p,n}$, we may lower bound $B_{p,t+1}$ as

$$B_{p,t+1} \geq \frac{w_p \sigma_p}{T_{p,t}} \geq \frac{w_p \sigma_p}{T_{p,n}} \ . \tag{21}$$

Combining Equations 19, 20, and 21, we obtain

$$\frac{w_p \sigma_p}{T_{p,n}} \leq \frac{w_k}{T_{k,n}-1} \left(\sigma_k + 4a \sqrt{\frac{\log(2/\delta)}{T_{k,n}-1}} \right) \ . \tag{22}$$

Note that at this point there is no dependency on t, and thus, the probability that Equation 22 holds for any p and for any k such that arm k is pulled after the initialization phase, i.e., such that $T_{k,n} > 2$, is at least $1 - 2nK\delta$ (probability of event ξ).

Step 2. Lower bound on $T_{p,n}$. If an arm p is under-pulled compared to its optimal allocation *without taking into account the initialization phase*,i.e., $T_{p,n} - 2 < \lambda_p(n - 2K)$, then from the constraint $\sum_k (T_{k,n} - 2) = n - 2K$ and the definition of the optimal allocation,

we deduce that there exists at least another arm k that is over-pulled compared to its optimal allocation *without taking into account the initialization phase*, i.e., $T_{k,n} - 2 > \lambda_k(n-2K)$. Note that for this arm, $T_{k,n} - 2 > \lambda_k(n-2K) \geq 0$, so we know that this specific arm is pulled at least once *after* the initialization phase and that it satisfies Equation 22. Using the definition of the optimal allocation $T^*_{k,n} = n w_k \sigma_k / \Sigma_w$, and the fact that $T_{k,n} \geq \lambda_k(n-2K)+2$, Equation 22 may be written as for any arm p

$$\begin{aligned}\frac{w_p\sigma_p}{T_{p,n}} &\leq \frac{w_k}{T^*_{k,n}} \frac{n}{(n-2K)} \left(\sigma_k + 4a\sqrt{\frac{\log(2/\delta)}{\lambda_k(n-2K)+1}} \right) \\ &\leq \frac{\Sigma_w}{n} + \frac{4K\Sigma_w}{n^2} + 8\sqrt{2}a \frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_k^{3/2}} \ ,\end{aligned}$$

because $n \geq 4K$. The previous Equation, combined with the fact that $\lambda_k \geq \lambda_{\min}$, may be written as

$$\frac{w_p\sigma_p}{T_{p,n}} \leq \frac{\Sigma_w}{n} + 12a \frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2} \ . \tag{23}$$

By rearranging Equation 23, we obtain the lower bound on $T_{p,n}$:

$$T_{p,n} \geq \frac{w_p\sigma_p}{\frac{\Sigma_w}{n} + 12a\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}} \geq T^*_{p,n} - 12a\lambda_p \frac{\sqrt{\log(2/\delta)}}{\Sigma_w \lambda_{\min}^{3/2}}\sqrt{n} - 4K\lambda_p \ , \tag{24}$$

where in the second inequality we use $1/(1+x) \geq 1-x$ (for $x > -1$). Note that the lower bound holds on ξ for any arm p.

Step 3. Upper bound on $T_{p,n}$. Using Equation 24 and the fact that $\sum_k T_{k,n} = n$, we obtain

$$T_{p,n} = n - \sum_{k\neq p} T_{k,n} \leq \left(n - \sum_{k\neq p} T^*_{k,n} \right) + \sum_{k\neq p} \left(12a\lambda_p \frac{\sqrt{\log(2/\delta)}}{\Sigma_w \lambda_{\min}^{3/2}}\sqrt{n} + 4K\lambda_p \right) .$$

And we deduce because $\sum_{k\neq p} \lambda_k \leq 1$

$$T_{p,n} \leq T^*_{p,n} + 12a \frac{\sqrt{\log(2/\delta)}}{\Sigma_w \lambda_{\min}^{3/2}}\sqrt{n} + 4K \ . \tag{25}$$

The lemma follows by combining the lower and upper bounds in Equations 24 and 25. ■

C.2 Proof of Theorem 2

We are now ready to prove Theorem 2.

Proof [Theorem 2] By definition, the pseudo-loss of the algorithm is

$$\begin{aligned}\mathbb{E}[L_n] &= \sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[\frac{\sigma_k^2}{T_{k,n}}\Big] = \sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[\frac{\sigma_k^2}{T_{k,n}}\mathbb{I}\{\xi\}\Big] + \sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[\frac{\sigma_k^2}{T_{k,n}}\mathbb{I}\{\xi^C\}\Big] \\ &\leq \sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2}{\underline{T}_{k,n}} + \sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2}{2}\mathbb{P}(\xi^c).\end{aligned}$$

where $\underline{T}_{k,n}$ is the lower bound on $T_{k,n}$ on the event ξ, and also because $T_{k,n} \geq 2$ by definition of algorithm MC-UCB.

Using Equation 23 for $w_k\sigma_k/\underline{T}_{k,n}$ (result of Lemma 15, which is equivalent to using a lower bound on $T_{k,n}$ on the event ξ), we obtain

$$\begin{aligned}\sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2}{\underline{T}_{k,n}} &\leq \sum_{k=1}^{K} w_k\sigma_k\Big(\frac{\Sigma_w}{n} + 12a\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}\Big) \\ &\leq \frac{\Sigma_w^2}{n} + 12a\Sigma_w\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w^2}{n^2}.\end{aligned}$$

Finally we have, because of Lemma 11 tells us that $\mathbb{P}(\xi^c) \leq 2nK\delta$, that

$$\begin{aligned}\mathbb{E}[L_n] &\leq \frac{\Sigma_w^2}{n} + 12a\Sigma_w\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w^2}{n^2} + \Sigma_{w,2}nK\delta \\ &\leq \frac{\Sigma_w^2}{n} + 168\sqrt{2c_1(c_2+2)\log(n)}\Sigma_w\frac{\sqrt{\log(n)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w^2}{n^2} + \frac{\Sigma_{w,2}}{n^{5/2}}K \\ &\leq \frac{\Sigma_w^2}{n} + 168\sqrt{2c_1(c_2+2)}\Sigma_w\frac{\log(n)}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{5K\Sigma_{w,2}}{n^2}.\end{aligned}$$

where we use $a \leq 2\sqrt{2c_1(c_2+2)\log(n)}$ and $\delta = n^{-7/2}$. Those bounds are made explicit in Appendix B.3.

This concludes the proof.

■

C.3 Proof of Proposition 4

We are also ready to prove Proposition 4

Proof [Proposition 4] The proof consists of the following two steps.

Step 1. $T_{k,n}$ is a stopping time. Consider an arm k. At each time step $t+1$, the MC-UCB algorithm decides which arm to pull according to the current values of the upper-bounds $\{B_{k,t+1}\}_k$. Thus for any arm k, $T_{k,(t+1)}$ depends only on the values $\{T_{k,t}\}_k$ and $\{\hat{\sigma}_{k,t}\}_k$. So by induction, $T_{k,(t+1)}$ depends on the sequence $\{X_{k,1}, \ldots, X_{k,T_{k,t}}\}$, and on the samples of the other arms (which are independent of the samples of arm k). We deduce that $T_{k,n}$ is a

stopping time adapted to the process $(X_{k,t})_{t\leq n}$.

Step 2. Bound on $\sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\Big]$. By definition, we have

$$\sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\Big] = \sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\Big] + \sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}\Big].$$

Using the definition of $\hat{\mu}_{k,n}$ and Proposition 13 we bound the first term as

$$\sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\Big] \leq \sum_{k=1}^K w_k^2\frac{\sigma_k^2\mathbb{E}[T_{k,n}]}{\underline{T}_{k,n}^2}\ , \tag{26}$$

where $\underline{T}_{k,n}$ is the lower bound on $T_{k,n}$ on the event ξ.

Note that as $\sum_k T_{k,n}=n$, we also have $\sum_k \mathbb{E}[T_{k,n}]=n$.

Using Equation 26 and Equation 23 for $w_k\sigma_k/\underline{T}_{k,n}$ (which is equivalent to using a lower bound on $T_{k,n}$ on the event ξ), we obtain

$$\sum_{k=1}^K w_k^2\frac{\sigma_k^2\mathbb{E}[T_{k,n}]}{\underline{T}_{k,n}^2} \leq \sum_{k=1}^K \Big(\frac{\Sigma_w}{n} + 12a\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}\Big)^2\mathbb{E}[T_{k,n}]. \tag{27}$$

Equation 27 may be bounded using the fact that $\sum_k \mathbb{E}[T_{k,n}]=n$ as

$$\begin{aligned}
\sum_{k=1}^K w_k^2\frac{\sigma_k^2\mathbb{E}[T_{k,n}]}{\underline{T}_{k,n}^2} &\leq \Big(\frac{\Sigma_w}{n} + 12a\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}\Big)^2 n \\
&\leq \Big((\frac{\Sigma_w}{n})^2 + 24a\Sigma_w\frac{\sqrt{\log(2/\delta)}}{n^{5/2}\lambda_{\min}^{3/2}} + \frac{8K\Sigma_w^2}{n^3} + 288a^2\frac{\log(2/\delta)}{n^3\lambda_{\min}^3} + \frac{8K^2\Sigma_w^2}{n^4}\Big)n \\
&= \frac{\Sigma_w^2}{n} + 24a\Sigma_w\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{8K\Sigma_w^2}{n^2} + 288a^2\frac{\log(2/\delta)}{n^2\lambda_{\min}^3} + \frac{8K^2\Sigma_w^2}{n^3} \\
&\leq \frac{\Sigma_w^2}{n} + 24a\Sigma_w\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{16}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 18a^2\log(2/\delta)\Big).
\end{aligned}$$

From Lemma 14, we have $\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}\Big] \leq 2c_1n^2K\delta(1+\log(c_2/2nK\delta))$. Thus using the previous equation, we deduce

$$\begin{aligned}
\sum_{k=1}^{K} w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\Big] \leq & \frac{\Sigma_w^2}{n} + 24a\Sigma_w\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{16}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 18a^2\log(2/\delta)\Big) \\
& + 2c_1n^2K\delta(1+\log(c_2/2nK\delta)) \\
\leq & \frac{\Sigma_w^2}{n} + 54a\Sigma_w\frac{\sqrt{\log(n)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{16}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 90a^2\log(n)\Big) \\
& + 6c_1K(c_2+1)\log(n)n^{-3/2} \\
\leq & \frac{\Sigma_w^2}{n} + \frac{\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}\Big(112\Sigma_w\sqrt{c_1(c_2+2)} + 6c_1(c_2+2)K\Big) \\
& + \frac{19}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 720c_1(c_2+1)\log(n)^2\Big)\,.
\end{aligned}$$

where we use $a \leq 2\sqrt{2c_1(c_2+2)\log(n)}$ and $\mathbb{E}\left[|\hat{\mu}_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\right] \leq 6c_1K(c_2+1)\log(n)n^{-3/2}$. Those bounds are made explicit in B.3.

The Theorem follows by expressing the regret.

■

Appendix D. Proof of Theorems 3 and Proposition 5

Again, we first state and prove the following Lemma and then use this result to prove Theorem 3 and Proposition 5.

D.1 Problem independent Bound on the number of pulls of each arm

Lemma 16 *Let Assumption 1 hold. For any $0 < \delta \leq 1$ and for $n \geq 4K$, the algorithm MC-UCB satisfies on ξ, and thus with probability at least $1-2nK\delta$, for any arm p,*

$$T_{p,n} \geq T^*_{p,n} - \Big(24aK^{1/3}\frac{1}{\Sigma_w}\lambda_q\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{2/3} + 12K\lambda_q\Big), \tag{28}$$

and

$$T_{p,n} \leq T^*_{p,n} + \Big(24aK^{1/3}\frac{1}{\Sigma_w}\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{2/3} + 12K\Sigma_w\Big), \tag{29}$$

where $c(\delta) = \Big(\frac{2a\sqrt{\log(2/\delta)}}{\Sigma_w+4a\sqrt{\log(2/\delta)}}\frac{1}{K}\Big)^{2/3}$ *and* $a = \sqrt{2c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}}n^{1/2}$.

Unlike the bounds proved in Lemma 15, the difference between $T_{p,n}$ and $T^*_{p,n}$ is bounded by $\widetilde{O}(n^{2/3})$ without any inverse dependency on $\lambda_{\min}$.

Proof [Proof of Lemma 16]

Step 1. Lower bound of order $\widetilde{O}(n^{2/3})$. Let k be the index of an arm that is such that $T_{k,n} - 2 \geq w_k(n - 2K)$ (this implies $T_{k,n} \geq 3$ as $n \geq 4K$, and arm k is thus pulled after the initialization)[14]. Let $t + 1 \leq n$ be the last time at which it was pulled, i.e., $T_{k,t} = T_{k,n} - 1$ and $T_{k,t+1} = T_{k,n}$. From Equation 15 and the fact that $T_{k,n} \geq w_k n$, we obtain on ξ

$$B_{k,t} \leq \frac{w_k}{T_{k,t}} \left(\sigma_k + 4a \sqrt{\frac{\log(2/\delta)}{T_{k,t}}} \right) \leq \frac{\left(\max_p \sigma_p + 4a \sqrt{\log(2/\delta)} \right)}{n}, \tag{30}$$

where the second inequality follows from the facts that $T_{k,t} \geq 1$, $w_k \sigma_k \leq \Sigma_w$, and $w_k \leq \sum_k w_k = 1$. Since at time $t + 1$ the arm k has been pulled, then for any arm q, we have

$$B_{q,t} \leq B_{k,t}. \tag{31}$$

From the definition of $B_{q,t}$, and also using the fact that $T_{q,t} \leq T_{q,n}$, we deduce on ξ that

$$B_{q,t} \geq 2a w_q \frac{\sqrt{\log(2/\delta)}}{T_{q,t}^{3/2}} \geq 2a w_q \frac{\sqrt{\log(2/\delta)}}{T_{q,n}^{3/2}} \,. \tag{32}$$

Combining Equations 30–32, we obtain on ξ

$$2a w_q \frac{\sqrt{\log(2/\delta)}}{T_{q,n}^{3/2}} \leq \frac{\max_p \sigma_p + 4a \sqrt{\log(2/\delta)}}{n}.$$

Finally, this implies on ξ that for any q,

$$T_{q,n} \geq \left(\frac{2a w_q \sqrt{\log(2/\delta)}}{\Sigma_w + 4a \sqrt{\log(2/\delta)}} n \right)^{2/3}. \tag{33}$$

In order to simplify the notation, in the following we define

$$c(\delta) = \left(\frac{2a \sqrt{\log(2/\delta)}}{\Sigma_w + 4a \sqrt{\log(2/\delta)}} \right)^{2/3},$$

thus the lower bound on $T_{q,n}$ on ξ writes $T_{q,n} \geq w_q^{2/3} c(\delta) n^{2/3}$.

Step 2. Properties of the algorithm. We follow a similar analysis to Step 1 of the proof of Lemma 15. We first recall the definition of $B_{q,t+1}$ used in the MC-UCB algorithm

$$B_{q,t+1} = \frac{w_q}{T_{q,t}} \left(\hat{\sigma}_{q,t} + 2a \sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right).$$

Using Corollary 12 it follows that, on ξ

$$\frac{w_q \sigma_q}{T_{q,t}} \leq B_{q,t+1} \leq \frac{w_q}{T_{q,t}} \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right). \tag{34}$$

14. Note that such an arm always exists for any possible allocation strategy, as $n - 2K = \sum_q (T_{q,n} - 2)$, $1 = \sum_q w_q$, and $\forall q, w_q > 0$.

Let $t+1 \geq 2K+1$ be the time at which an arm q is pulled for the last time, that is $T_{q,t} = T_{q,n} - 1$. Note that there is at least one arm such that this happens as $n \geq 4K$. Since at $t+1$ arm q is chosen, then for any other arm p, we have

$$B_{p,t+1} \leq B_{q,t+1} \,. \tag{35}$$

From Equation 34 and $T_{q,t} = T_{q,n} - 1$, we obtain on ξ

$$B_{q,t+1} \leq \frac{w_q}{T_{q,t}} \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right) = \frac{w_q}{T_{q,n}-1} \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}} \right). \tag{36}$$

Furthermore, since $T_{p,t} \leq T_{p,n}$, then on ξ

$$B_{p,t+1} \geq \frac{w_p \sigma_p}{T_{p,t}} \geq \frac{w_p \sigma_p}{T_{p,n}}. \tag{37}$$

Combining Equations 35–37, we obtain on ξ

$$\frac{w_p \sigma_p}{T_{p,n}} (T_{q,n} - 1) \leq w_q \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}} \right).$$

Summing over all q such that the previous Equation is verified, i.e. such that $T_{q,n} \geq 3$, on both sides, we obtain on ξ

$$\frac{w_p \sigma_p}{T_{p,n}} \sum_{q | T_{q,n} \geq 3} (T_{q,n} - 1) \leq \sum_{q | T_{q,n} \geq 3} w_q \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}} \right).$$

This implies

$$\frac{w_p \sigma_p}{T_{p,n}} (n - 2K) \leq \sum_{q=1}^{K} w_q \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}} \right). \tag{38}$$

Step 3. Lower bound. Plugging Equation 33 in Equation 38,

$$\begin{aligned}
\frac{w_p \sigma_p}{T_{p,n}} (n - 2K) &\leq \sum_q w_q \left(\sigma_q + 4a \sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}} \right) \\
&\leq \sum_q w_q \left(\sigma_q + 4a \sqrt{\frac{2\log(2/\delta)}{w_q^{2/3} c(\delta) n^{2/3}}} \right) \\
&\leq \Sigma_w + \sum_q 4a w_q^{2/3} \sqrt{2 \frac{\log(2/\delta)}{c(\delta) n^{2/3}}} \leq \Sigma_w + 6a K^{1/3} \sqrt{\frac{\log(2/\delta)}{c(\delta) n^{2/3}}},
\end{aligned}$$

on ξ, since $\sum_q w_q^{2/3} \leq K^{1/3}$ by Jensen inequality and because $T_{q,n} - 1 \geq \frac{T_{q,n}}{2}$ (as $T_{q,n} \geq 2$). Finally as $n \geq 4K$, we obtain on ξ the following bound

$$\frac{w_p \sigma_p}{T_{p,n}} \leq \frac{\Sigma_w}{n} + 24 a K^{1/3} \sqrt{\frac{\log(2/\delta)}{c(\delta)}} n^{-4/3} + \frac{12 K \Sigma_w}{n^2}. \tag{39}$$

We now invert the bound and obtain on ξ the final lower-bound on $T_{p,n}$ as follows:

$$T_{p,n} \geq \frac{w_p \sigma_p}{\frac{\Sigma_w}{n} + 24aK^{1/3}\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{-4/3} + \frac{12K\Sigma_w}{n^2}} \geq T^*_{p,n} - 24aK^{1/3}\frac{1}{\Sigma_w}\lambda_p\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{2/3} - 12K\lambda_p,$$

as $\frac{1}{1+x} \geq 1-x$. Note that the above lower bound holds with high probability for any arm p.

Step 4. Upper bound. An upper bound on $T_{p,n}$ on ξ follows by using $T_{p,n} = n - \sum_{q\neq p} T_{q,n}$ and the previous lower bound, that is

$$\begin{aligned} T_{p,n} &\leq n - \sum_{q\neq p} T^*_{q,n} + \sum_{q\neq p}\Big(12K\lambda_q + 24aK^{1/3}\frac{1}{\Sigma_w}\lambda_q\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{2/3}\Big) \\ &\leq T^*_{p,n} + \Big(24aK^{1/3}\frac{1}{\Sigma_w}\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{2/3} + 12K\Big), \end{aligned}$$

because $\sum_{q\neq p}\lambda_q \leq 1$. ■

D.2 Proof of Theorem 3

We are now ready to prove Theorem 3.

Proof [Theorem 3]

By definition, the pseudo-loss of the algorithm is

$$\begin{aligned} \mathbb{E}[L_n] &= \sum_{k=1}^K w_k^2 \mathbb{E}\Big[\frac{\sigma_k^2}{T_{k,n}}\Big] = \sum_{k=1}^K w_k^2 \mathbb{E}\Big[\frac{\sigma_k^2}{T_{k,n}}\mathbb{I}\{\xi\}\Big] + \sum_{k=1}^K w_k^2 \mathbb{E}\Big[\frac{\sigma_k^2}{T_{k,n}}\mathbb{I}\{\xi^C\}\Big] \\ &\leq \sum_{k=1}^K w_k^2 \frac{\sigma_k^2}{\underline{T}_{k,n}} + \sum_{k=1}^K w_k^2 \frac{\sigma_k^2}{2}\mathbb{P}(\xi^c). \end{aligned}$$

where $\underline{T}_{k,n}$ is the lower bound on $T_{k,n}$ on the event ξ, and also because $T_{k,n} \geq 2$ by definition of algorithm MC-UCB.

Using Equation 39 for $w_k\sigma_k/\underline{T}_{k,n}$ (result of Lemma 16, which is equivalent to using a lower bound on $T_{k,n}$ on the event ξ), we obtain

$$\begin{aligned} \sum_{k=1}^K w_k^2 \frac{\sigma_k^2}{\underline{T}_{k,n}} &\leq \sum_{k=1}^K w_k\sigma_k\Big(\frac{\Sigma_w}{n} + 24aK^{1/3}\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{-4/3} + \frac{12K\Sigma_w}{n^2}\Big) \\ &\leq \frac{\Sigma_w^2}{n} + 24aK^{1/3}\Sigma_w\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{-4/3} + \frac{12K\Sigma_w^2}{n^2}. \end{aligned} \tag{40}$$

Finally we have, as by Lemma 11, we know that $\mathbb{P}(\xi^c) \leq 2nK\delta$, that

$$\begin{aligned}\mathbb{E}[L_n] &\leq \frac{\Sigma_w^2}{n} + 24aK^{1/3}\Sigma_w\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{-4/3} + \frac{12K\Sigma_w^2}{n^2} + \Sigma_{w,2}nK\delta \\ &\leq \frac{\Sigma_w^2}{n} + 336\sqrt{2c_1(c_2+2)}(\sqrt{c_2}+1)^{2/3}K^{1/3}\Sigma_w\frac{\log(n)}{n^{4/3}} + \frac{5K\Sigma_{w,2}}{n^2},\end{aligned}$$

where we use $a \leq 2\sqrt{2c_1(c_2+2)\log(n)}$, $c(\delta) \geq \left(\frac{1}{\sqrt{c_2}+1}\right)^{2/3}$ and $\delta = n^{-7/2}$. These bounds are made explicit in Appendix B.3.

This concludes the proof.

■

D.3 Proof of Proposition 5

We are also ready to prove Proposition 5.

Proof [Proposition 5]

We decompose $\sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\Big]$ on ξ and its complement:

$$\sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\Big] = \sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\Big] + \sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}\Big].$$

Using the definition of $\hat{\mu}_{k,n}$ and Proposition 13 we bound the first term as

$$\sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\Big] \leq \sum_{k=1}^K w_k^2\frac{\sigma_k^2\mathbb{E}[T_{k,n}]}{\underline{T}_{k,n}^2}\,, \tag{41}$$

where $\underline{T}_{k,n}$ is the lower bound on $T_{k,n}$ on ξ.

Note also that as $\sum_k T_{k,n} = n$, we also have $\sum_k \mathbb{E}[T_{k,n}] = n$. Using Equation 41 and Equation 39 which provides an upper bound on ξ on $\frac{w_k\sigma_k}{T_{k,n}}$ (and thus a lower bound on ξ on $T_{k,n}$), we deduce

$$\sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\left\{\xi\right\}\Big] \leq \sum_{k=1}^K \Big(\frac{\Sigma_w}{n} + 24aK^{2/3}\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{-4/3} + \frac{12K\Sigma_w}{n^2}\Big)^2\mathbb{E}[T_{k,n}]. \tag{42}$$

Using the fact that $\sum_k \mathbb{E}[T_{k,n}] = n$, Equation 42 may be rewritten as

$$\begin{aligned}
\sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2 \mathbb{I}\left\{\xi\right\}\Big] \leq & \Big(\frac{\Sigma_w}{n} + 24aK^{2/3}\sqrt{\frac{\log(2/\delta)}{c(\delta)}}n^{-4/3} + \frac{12K\Sigma_w}{n^2}\Big)^2 n \\
\leq & \Big((\frac{\Sigma_w}{n})^2 + \frac{48\Sigma_w a K^{2/3}}{n^{7/3}}\sqrt{\frac{\log(2/\delta)}{c(\delta)}} \\
& + \frac{12K\Sigma_w^2}{n^3} + \frac{1152a^2K^{4/3}}{n^{8/3}}\frac{\log(2/\delta)}{c(\delta)} + \frac{288K^2\Sigma_w^2}{n^4}\Big)n \\
= & \frac{\Sigma_w^2}{n} + \frac{48\Sigma_w a K^{2/3}}{n^{4/3}}\sqrt{\frac{\log(2/\delta)}{c(\delta)}} \\
& + \frac{12K\Sigma_w^2}{n^2} + \frac{1152a^2K^{4/3}}{n^{5/3}}\frac{\log(2/\delta)}{c(\delta)} + \frac{288K^2\Sigma_w^2}{n^3} \\
\leq & \frac{\Sigma_w^2}{n} + \frac{48\Sigma_w a K^{2/3}}{n^{4/3}}\sqrt{\frac{\log(2/\delta)}{c(\delta)}} + \frac{300}{n^2}\Big(4a^2K^{4/3}\frac{\log(2/\delta)}{c(\delta)} + K\Sigma_w^2\Big).
\end{aligned}$$

From Lemma 14, we have $\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}\Big] \leq 2c_1 n^2 K\delta(1+\log(c_2/2nK\delta))$. Thus using the last equation and the fact that $\delta = n^{-7/2}$, the loss is bounded as

$$\begin{aligned}
& \sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\Big] \\
\leq & \frac{\Sigma_w^2}{n} + \frac{48\Sigma_w a K^{2/3}}{n^{4/3}}\sqrt{\frac{\log(2/\delta)}{c(\delta)}} + \frac{300}{n^2}\Big(4a^2K^{4/3}\frac{\log(2/\delta)}{c(\delta)} + K\Sigma_w^2\Big) + 2c_1 n^2 K\delta(1+\log(c_2/2nK\delta)) \\
\leq & \frac{\Sigma_w^2}{n} + \frac{96\Sigma_w aK}{n^{4/3}}\sqrt{\log(n)}\big(\sqrt{c_2}+1\big)^{1/3} + \frac{300}{n^2}\Big(16a^2K^2\log(n)\big(\sqrt{c_2}+1\big)^{2/3} + K\Sigma_w^2\Big) \\
& + 6c_1K(c_2+1)\log(n)n^{-3/2} \\
\leq & \frac{\Sigma_w^2}{n} + \frac{200\sqrt{c_1(c_2+2)}\Sigma_w K}{n^{4/3}}\log(n)\big(\sqrt{c_2}+1\big)^{1/3} \\
& + \frac{365}{n^{3/2}}\Big(16a^2K^2\log(n)\big(\sqrt{c_2}+1\big)^{2/3} + K\Sigma_w^2 + c_1(c_2+2)K\log(n)\Big) \\
\leq & \frac{\Sigma_w^2}{n} + \frac{200\sqrt{c_1}(c_2+2)\Sigma_w K}{n^{4/3}}\log(n) + \frac{365}{n^{3/2}}\Big(129c_1(c_2+2)^2K^2\log(n)^2 + K\Sigma_w^2\Big).
\end{aligned}$$

where we use $a \leq 2\sqrt{2c_1(c_2+2)\log(n)}$, $c(\delta) \geq \Big(\frac{1}{\sqrt{c_2}+1}\Big)^{2/3}$ and $\mathbb{E}\big[|\hat{\mu}_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq 6c_1K(c_2+1)\log(n)n^{-3/2}$. Those bound are made explicit in B.3.

■

Appendix E. Comments on problem independent bound for GAFS-WL

Let $n \geq 4$ be the budget. We face a two-arms bandit problem with $w_1 = w_2 = \frac{1}{2}$ and such that (i) the distribution of the first arm is a Bernoulli of parameter $p = \frac{1}{n^{1/2+\epsilon}}$ with ϵ such that $1/6 > \epsilon > 0$ and that (ii) the distribution of the second arm is such that $\sigma_2 = 1$ and bounded by c.

Note that

$$\frac{1}{2n^{1/4+\epsilon/2}} \leq \sigma_1 \leq \frac{1}{n^{1/4+\epsilon/2}} \qquad \text{and} \qquad \sigma_2 = 1,$$

because $\sigma_1 = \sqrt{p(1-p)}$ and that thus

$$L_n^* \leq \frac{(1+n^{-1/4-\epsilon/2})^2}{4n} \leq \frac{1+3n^{-1/4-\epsilon/2}}{4n} \leq \frac{1}{4n} + \frac{1}{n^{5/4+\epsilon/2}}.$$

We run algorithm GAFS-WL on that problem. Note that algorithm GAFS-WL pull each arm $\lfloor a\sqrt{n} \rfloor$ times and then pull the arms according to $\frac{w_k \hat{\sigma}_{k,t}}{T_{k,t}}$.

We call $\{X_{p,u}\}_{p=1,2;u=1,\dots,n}$ the samples of the arms.

Note that:

$$\begin{aligned}
\mathbb{P}\Big(X_{1,1} = 0, \dots, X_{1,\lfloor a\sqrt{n} \rfloor} = 0\Big) &\geq (1 - \frac{1}{n^{1/2+\epsilon}})^{a\sqrt{n}} \\
&\geq (1 - \frac{an^{-\epsilon}}{a\sqrt{n}})^{a\sqrt{n}} \\
&\geq (1 - an^{-\epsilon}) \exp(-an^{-\epsilon}) \geq (1 - an^{-\epsilon})^2.
\end{aligned}$$

Note on the other hand, that $\mathbb{P}(|\hat{\sigma}_{2,a\sqrt{n}} - 1| \geq \frac{2\sqrt{\log(2/\delta)}}{\sqrt{a}n^{1/4}}) \leq \delta$. This means that with probability at least $1 - 2\exp(-a\sqrt{n}/4)$, we have $\hat{\sigma}_{2,a\sqrt{n}} > 0$.

The probability that $\hat{\sigma}_{1,a\sqrt{n}} = 0$ goes to 1 when n goes to $+\infty$. The probability that $\hat{\sigma}_{2,a\sqrt{n}} > 0$ goes to 1 when n goes to $+\infty$. This means that the probability that GAFS-WL stops pulling arm 1 after $a\sqrt{n}$ pulls goes to 1 when n goes to $+\infty$, and arm 1 is under-pulled if $\epsilon < 1/2$ (it should be pulled $n^{3/4-\epsilon/2}$).

Note that on the event such that $\Big(X_{1,1} = 0, \dots, X_{1,\lfloor a\sqrt{n} \rfloor} = 0\Big)$, we know that $\hat{\mu}_{1,a\sqrt{n}} = 0$. Note also that we know that as arm 2 is gaussian, we have $\mathbb{E}(\hat{\mu}_{2,n} - \mu_2)^2 \leq \frac{1}{4n}$. The performance of GAFS-WL then verifies

$$\begin{aligned}
\mathbb{E}\Big[\sum_k w_k^2 (\hat{\mu}_{k,n} - \mu_k)^2\Big] &\geq \frac{1}{4n} + \mathbb{P}(\hat{\sigma}_{1,a\sqrt{n}} = 0)\mathbb{P}(\hat{\sigma}_{2,a\sqrt{n}} > 0)\Big(n^{-1/2-\epsilon}\Big)^2 \\
&\geq \frac{1}{4n} + (1 - 2\exp(-a\sqrt{n}/4))(1 - an^{-\epsilon})^2\Big(n^{-1-2\epsilon}\Big) \\
&\geq \frac{1}{4n} + (1 - \frac{8}{a\sqrt{n}})(1 - 2\frac{a}{n^{\epsilon}})\frac{1}{n^{1+2\epsilon}} \\
&\geq \frac{1}{4n} + \frac{1}{n^{1+2\epsilon}} - \frac{8}{an^{3/2+2\epsilon}} - \frac{2a}{n^{1+3\epsilon}} \\
&\geq \frac{1}{4n} + \frac{1}{n^{1+2\epsilon}} - \frac{10\max(a, 1/a)}{n^{1+3\epsilon}},
\end{aligned}$$

where the last line is obtained using the fact that $\epsilon < 1/6$. Note that we used the proxy defined in paper Grover (2009) to measure performance, so that we can compare with their bound.

We thus have

$$\mathbb{E}\Big[\sum_k w_k^2(\hat{\mu}_{k,n}-\mu_k)^2\Big] - \frac{\Sigma_w^2}{n} \geq \frac{1}{n^{1+2\epsilon}} - \frac{10\max(a,1/a)}{n^{1+3\epsilon}} - \frac{1}{n^{5/4+\epsilon/2}}$$
$$\geq \frac{1}{n^{1+2\epsilon}} - \frac{11\max(a,1/a)}{n^{1+3\epsilon}},$$

again because $\epsilon < 1/6$. This implies that for n such that $n \geq (\frac{11\max(a,1/a)}{2})^{1/\epsilon}$, we have

$$\mathbb{E}\Big[\sum_k w_k^2(\hat{\mu}_{k,n}-\mu_k)^2\Big] - \frac{\Sigma_w^2}{n} \geq \frac{1}{2n^{1+2\epsilon}},$$

with ϵ arbitrarily close to 0.

Appendix F. Proof of Propositions 6, 7 and 8

F.1 Proof of Proposition 6

We first prove that the bounds of Theorems 4 and 5 can be directly expressed as bounds on the mean squared error $\mathbb{E}[(\hat{\mu}_n - \mu)^2]$ when the distributions of the arms are symmetric.

Proof [Proof of Proposition 6]

Step 1: Expression of $\mathbb{E}[\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)|T_{k,n}=T_1,T_{q,n}=T_2]$. At each time step $t+1 > 2K$, the MC-UCB algorithm decides which arm to pull according to the current values of the upper-bounds $\{B_{p,t+1}\}_p$. Thus for any arm k, $T_{k,(t+1)}$ depends only of the values $\{T_{p,t}\}_p$ and $\{\hat{\sigma}_{p,t}\}_p$. So by induction, $T_{k,n}$ depends of the samples of the arms only trough the K sequences $\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}$.

Let us consider another arm $q \neq k$. The samples of arm k and arm q depend of each other only trough $(T_{k,t})_{t\leq n}$ and $(T_{q,t})_{t\leq n}$, and thus by induction only trough the sequence $\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}$. The samples are thus independent conditionally to the $\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}$.

This leads to:

$$\begin{aligned}
&\mathbb{E}[\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)|T_{k,n}=T_1,T_{q,n}=T_2]\\
&=\mathbb{E}\Big[\big(\frac{1}{T_1}\sum_{u=1}^{T_1}X_{k,u}-\mu_k\big)\big(\frac{1}{T_2}\sum_{u=1}^{T_2}X_{q,u}-\mu_q\big)|T_{k,n}=T_1,T_{q,n}=T_2\Big]\\
&=\mathbb{E}\Big[\mathbb{E}\big[\big(\frac{1}{T_1}\sum_{u=1}^{T_1}X_{k,u}-\mu_k\big)\big(\frac{1}{T_2}\sum_{u=1}^{T_2}X_{q,u}-\mu_q\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big]\\
&\times\mathbb{P}\big(\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}|T_{k,n}=T_1,T_{q,n}=T_2\big)|T_{k,n}=T_1,T_{q,n}=T_2\Big]\\
&=\mathbb{E}\Big[\mathbb{E}\big[\big(\frac{1}{T_1}\sum_{u=1}^{T_1}X_{k,u}-\mu_k\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big]\mathbb{P}\big(\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}|T_{k,n}=T_1,T_{q,n}=T_2\big)|T_{k,n}=T_1,T_{q,n}=T_2\Big]\\
&\times\mathbb{E}\Big[\mathbb{E}\big[\big(\frac{1}{T_2}\sum_{u=1}^{T_2}X_{q,u}-\mu_q\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big]\mathbb{P}\big(\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}|T_{k,n}=T_1,T_{q,n}=T_2\big)|T_{k,n}=T_1,T_{q,n}=T_2\Big],
\end{aligned}\tag{43}$$

where the $X_{p,u}$ are the u-th samples pulled from arm p.

Step 2: The distribution of $\sum_{u=1}^{T}X_{k,u}-\mu_k$ conditioned on $\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}$ is symmetric. Consider an arm k, and a time T. As the distributions ν_k is symmetric, $\frac{1}{T}\sum_{u=1}^{T}X_{k,u}-\mu_k$ conditioned on $\{\hat{\sigma}_{k,t'}\}_{t'\leq n}$ is symmetric.

As $\frac{1}{T}\sum_{u=1}^{T}X_{k,u}-\mu_k$ depends on $\{\hat{\sigma}_{p,t'}\}_{p\neq k,t'\leq n}$ only trough $\{\hat{\sigma}_{k,t'}\}_{t'\leq n}$, the $\frac{1}{T}\sum_{u=1}^{T}X_{k,u}-\mu_k$ conditioned on $\{\hat{\sigma}_{k,t'}\}_{t'\leq n}$ is independent of $\{\hat{\sigma}_{p,t'}\}_{p\neq k,t'\leq n}$. The distribution of $\frac{1}{T}\sum_{u=1}^{T}X_{k,u}-\mu_k$ conditioned on $\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}$ is thus symmetric around 0, as ν_k is symmetric around μ_k.

This leads to

$$\mathbb{E}\big[\big(\frac{1}{T}\sum_{u=1}^{T}X_{k,u}-\mu_k\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big]=0.\tag{44}$$

Step 4: The cross products $\mathbb{E}[\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)]$ are null. We combine Equations 43 and 44 to get

$$\begin{aligned}
&\mathbb{E}[\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)|T_{k,n}=T_1,T_{q,n}=T_2]\\
&=\mathbb{E}\Big[0|T_{k,n}=T_1,T_{q,n}=T_2\Big]\mathbb{E}\Big[0|T_{k,n}=T_1,T_{q,n}=T_2\Big]=0,
\end{aligned}$$

Now note that

$$\begin{aligned}
&\mathbb{E}\Big[\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)\Big]\\
&=\sum_{T_1=2}^{n}\sum_{T_2=2}^{n}\mathbb{E}\Big[\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)|T_{k,n}=T_1,T_{q,n}=T_2\Big]\mathbb{P}\big(T_{k,n}=T_1,T_{q,n}=T_2\big)=0,
\end{aligned}$$

where we use the previous Equation at the end.

Finally, we conclude the proof with

$$\begin{aligned}\mathbb{E}\Big[(\hat{\mu}_n-\mu)^2\Big] &= \mathbb{E}\Big[\big(\sum_{k=1}^{K} w_k(\hat{\mu}_{k,n}-\mu_k)\big)^2\Big] \\ &= \sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\Big] + 2\sum_{k\neq q} w_k w_q \mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)(\hat{\mu}_{q,n}-\mu_q)\Big] \\ &= L_n(\mathcal{A}_{MC-UCB}).\end{aligned}$$

■

F.2 Proof of Propositions 7 and 8

We also relate the bounds in Theorems 4 and 5 to a bound on $\mathbb{E}[(\hat{\mu}_n-\mu)^2]$ in the general case. The proof Propositions 7 and 8 are very similar up to the end, where we use for the problem dependent Proposition 7 the results of Lemma 15, and for the problem independent Proposition 8 the results of Lemma 16.

Proof

Step 0: A useful Lemma.

Lemma 17 *Let X be a random variables such that $\mathbb{E}(X)=0$. Let $(\Omega_u)_{u=1,\dots,p}$ be a partition of the space of random events. Let $(a_u)_{u=1,\dots,p}$ be a positive decreasing sequence of random numbers. We have*

$$|\mathbb{E}(X\sum_{u=1}^{p} a_u \mathbb{I}\{X\in\Omega_u\})| \leq (a_1-a_p)\sqrt{\mathbb{E}(X^2)}.$$

Proof

First note that as the sequence of a_u is positive decreasing, the following equation holds

$$X\sum_{u=1}^{p} a_u \mathbb{I}\{X\in\Omega_u\} \leq Xa_1\mathbb{I}\{X\geq 0\} + Xa_p\mathbb{I}\{X<0\}.$$

This implies

$$\begin{aligned}\mathbb{E}\Big[X\sum_{u=1}^{p} a_u \mathbb{I}\{X\in\Omega_u\}\Big] &\leq \mathbb{E}\Big[Xa_1\mathbb{I}\{X\geq 0\} + Xa_p\mathbb{I}\{X<0\}\Big] \\ &\leq \mathbb{E}\Big[(a_1-a_p)X\mathbb{I}\{X\geq 0\} + a_pX(\mathbb{I}\{X<0\}+\mathbb{I}\{X\geq 0\})\Big] \\ &\leq (a_1-a_p)\mathbb{E}\Big[X\mathbb{I}\{X\geq 0\}\Big] \\ &\leq (a_1-a_p)\sqrt{\mathbb{E}\Big[X^2\mathbb{I}\{X\geq 0\}\Big]} \\ &\leq (a_1-a_p)\sqrt{\mathbb{E}\Big[X^2\Big]},\end{aligned}$$

where the fourth line follows by Cauchy-Schwartz.

By remarking that

$$X\sum_{u=1}^{p} a_u \mathbb{I}\left\{X \in \Omega_u\right\} \geq X a_1 \mathbb{I}\left\{X \leq 0\right\} + X a_p \mathbb{I}\left\{X > 0\right\},$$

we prove in the same way that

$$\mathbb{E}\Big[X\sum_{u=1}^{p} a_u \mathbb{I}\left\{X \in \Omega_u\right\}\Big] \geq -(a_1 - a_p)\sqrt{\mathbb{E}\Big[X^2\Big]}.$$

Those two inequalities lead to the desired result.

■

Note first that

$$\mathbb{E}[(\hat{\mu}_n - \mu)^2] = \sum_{k\neq q} w_k^2 \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)^2\Big] + 2\sum_{k\neq q} w_k w_q \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\Big].$$

As problem dependent and problem independent bounds on $\sum_{k\neq q} w_k^2 \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)^2\Big]$ are available in Theorem 4 and 5, it is sufficient to bound the cross-products.

Step 1: $\mathbb{E}\Big[\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)\big(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\big)\Big] = 0$**.** Let us denote by $t_{k,t}$ the moment where the algorithm pulls arm k the $t-$th time.

$$\begin{aligned}
&\mathbb{E}\Big[\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)\big(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\big)\Big] \\
&= \mathbb{E}\Big[\big(\sum_{t=1}^{n}(X_{k,t}-\mu_k)\mathbb{I}\left\{T_{k,n}\geq t\right\}\big)\big(\sum_{t=1}^{n}(X_{q,t}-\mu_q)\mathbb{I}\left\{T_{q,n}\geq t\right\}\big)\Big] \\
&= \sum_{t=1}^{n}\sum_{t'=1}^{n}\mathbb{E}\Big[\big(X_{k,t}-\mu_k\big)\big(X_{q,t'}-\mu_q\big)\mathbb{I}\left\{T_{q,n}\geq t'\right\}\mathbb{I}\left\{T_{k,n}\geq t\right\}\Big] \\
&= \sum_{t=1}^{n}\sum_{t'=1}^{n}\mathbb{E}\Big[\big(X_{k,t}-\mu_k\big)\big(X_{q,t'}-\mu_q\big)\mathbb{I}\left\{T_{q,n}\geq t'\right\}\mathbb{I}\left\{T_{k,n}\geq t\right\}\mathbb{I}\left\{t_{k,t}<t_{q,t'}\right\}\Big] \\
&+ \sum_{t=1}^{n}\sum_{t'=1}^{n}\mathbb{E}\Big[\big(X_{k,t}-\mu_k\big)\big(X_{q,t'}-\mu_q\big)\mathbb{I}\left\{T_{q,n}\geq t'\right\}\mathbb{I}\left\{T_{k,n}\geq t\right\}\mathbb{I}\left\{t_{k,t}>t_{q,t'}\right\}\Big].
\end{aligned}$$

Let us call $\mathcal{F}_{t_1,\dots,t_K} = \sigma\Big(X_{1,1},\dots,X_{1,t_1},\dots,X_{K,1},\dots,X_{K,t_K}\Big)$ the multidimensional filtration generated, for all k, by the t_k first instance of the $k-$th arm. Note that the algorithm MC-UCB disposes at time t of the informations from a certain $\mathcal{F}_{t_1,\dots,t_K}$ where $\sum_k t_k = t$ and picks an arm (i.e. a dimension of the filtration) according *only* to information in $\mathcal{F}_{t_1,\dots,t_K}$. If the algorithm picks arm k, the information at the disposal of MC-UCB is, after pulling arm k, in $\mathcal{F}_{t_1,\dots,t_k+1,\dots,t_K}$.

Now let us consider consider two arms k and q. Note that the collection of events $\tau = \sigma(X_{q,t'}) \cap \{T_{q,n} \geq t'\} \cap \{T_{k,n} \geq t\} \cap \{t_{k,t} > t_{q,t'}\}$ is in $\mathcal{F}_{n,\dots,t-1,\dots,n}$[15]: indeed, no information of $X_{k,u}$ with u greater than $t-1$ is needed in addition $\mathcal{F}_{n,\dots,t-1,\dots,n}$ to know if we are in an event of τ and in which one. This means that $X_{k,t}$ is independent of all events in τ. Finally, we have

$$\begin{aligned}
&\mathbb{E}\Big[\big(X_{k,t}-\mu_k\big)\big(X_{q,t'}-\mu_q\big)\mathbb{I}\left\{T_{q,n}\geq t'\right\}\mathbb{I}\left\{T_{k,n}\geq t\right\}\mathbb{I}\left\{t_{k,t}>t_{q,t'}\right\}\Big] \\
&= \mathbb{E}\Big[\big(X_{q,t'}-\mu_q\big)\mathbb{I}\left\{T_{q,n}\geq t'\right\}\mathbb{I}\left\{T_{k,n}\geq t\right\}\mathbb{I}\left\{t_{k,t}\leq t_{q,t'}\right\}\mathbb{E}\big[\big(X_{k,t}-\mu_k\big)|\mathcal{F}_{n,\dots,t-1,\dots,n}\big]\Big] \\
&= \mathbb{E}\Big[\big(X_{q,t'}-\mu_q\big)\mathbb{I}\left\{T_{q,n}\geq t'\right\}\mathbb{I}\left\{T_{k,n}\geq t\right\}\mathbb{I}\left\{t_{k,t}>t_{q,t'}\right\}0\Big] = 0.
\end{aligned}$$

By summing and doing the same reasoning for arm q, we obtain that

$$\mathbb{E}\Big[\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)\big(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\big)\Big] = 0. \tag{45}$$

Note that we have by doing a similar reasoning, that

$$\mathbb{E}\Big[\big(\sum_{t=\max(T_{k,n},\underline{T}_k)}^{\min(T_{k,n},\bar{T}_k)}(X_{k,t}-\mu_k)\big)\big(\sum_{t'=\max(T_{q,n},\underline{T}_q)}^{\min(T_{q,n},\bar{T}_q)}(X_{q,t'}-\mu_q)\big)\Big] = 0, \tag{46}$$

where $\underline{T}_k$, $\underline{T}_q$, $\bar{T}_k$ and $\bar{T}_q$ are *any constants*.

Step 2: Definition of an event τ of high probability. We remind that on ξ, by combining Lemmas 15 and 16, we have for all p,

$$T_{p,n} \geq \underline{T}_{p,n} = \max\Big(T^*_{p,n} - B\sqrt{n}, T^*_{p,n} - A\lambda_p n^{2/3}, En^{2/3}\Big),$$

and

$$T_{p,n} \leq \bar{T}_{p,n} = \min\Big(T^*_{p,n} + D\sqrt{n}, T^*_{p,n} + Cn^{2/3}\Big),$$

where B and D are as in Lemma 15, A and C are as in Lemma 16, and E is as in the proof of Lemma 16 (Equation 33). Note that B and D display an invert dependency in $\lambda_{\min}$, but that A, C, and E do not. The probability of ξ is more than $1-2nK\delta$.

Now let us define the event τ such that for all p,

$$T_{p,n} \geq \underline{T}_{p,n} = \max\Big(T^*_{p,n} - B\sqrt{n}, T^*_{p,n} - A\lambda_p n^{2/3}, En^{2/3}\Big),$$

and

$$T_{p,n} \leq \bar{T}_{p,n} = \min\Big(T^*_{p,n} + D\sqrt{n}, T^*_{p,n} + Cn^{2/3}\Big).$$

15. Here there are n at all positions except at the $k-1$ where there is a t.

Note that $\xi \subset \tau$ because of Lemmas 15 and 16. We have, because of $\xi \subset \tau$,

$$
\begin{aligned}
&|\mathbb{E}[(\hat{\mu}_{q,n}-\mu_q)(\mu_{k,n}-\mu_k)\mathbb{I}\{\tau^c\}]| \qquad (47)\\
&\leq \sqrt{\mathbb{E}[(\hat{\mu}_{q,n}-\mu_q)^2\mathbb{I}\{\tau^c\}]}\sqrt{\mathbb{E}[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\tau^c\}]}\\
&\leq \sqrt{\mathbb{E}[(\hat{\mu}_{q,n}-\mu_q)^2\mathbb{I}\{\xi^c\}]}\sqrt{\mathbb{E}[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi^c\}]}\\
&\leq 2c_1 n^2 K\delta(1+\log(c_2/2nK\delta))\\
&\leq 2c_1 K(1+\log(c_2 n^{5/2}/2K))n^{-3/2}\\
&\leq C_\tau n^{-3/2}, \qquad (48)
\end{aligned}
$$

as in Appendix B and because $\delta = n^{-7/2}$. Here $C_\tau = 2c_1K(1+\log(c_2n^{5/2}/2K))$.

Step 3: Bounding the cross-products. Using step 1 and 2 together, we get

$$
\begin{aligned}
&\mathbb{E}\Big[\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)\big(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\big)\mathbb{I}\{\tau\}\Big]\\
&=\mathbb{E}\Big[\big(\sum_{t=\max(T_{k,n},\underline{T}_{k,n})}^{\min(T_{k,n},\bar{T}_{k,n})}(X_{k,t}-\mu_k)\big)\big(\sum_{t'=\max(T_{q,n},\underline{T}_{q,n})}^{\min(T_{q,n},\bar{T}_{q,n})}(X_{q,t'}-\mu_q)\big)\Big]=0.
\end{aligned}
$$

Let us call $Z = \big(\sum_{t=\max(T_{k,n},\underline{T}_{k,n})}^{\min(T_{k,n},\bar{T}_{k,n})}(X_{k,t}-\mu_k)\big)\big(\sum_{t'=\max(T_{q,n},\underline{T}_{q,n})}^{\min(T_{q,n},\bar{T}_{q,n})}(X_{q,t'}-\mu_q)\big)$. Note that $\mathbb{E}[Z]=0$. We thus have by Lemma 17

$$
\begin{aligned}
&\Big|\mathbb{E}\Big[\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)\mathbb{I}\{\tau\}\Big]\Big|\\
&=\Big|\mathbb{E}\Big[\big(\frac{1}{T_{k,n}}\sum_{t=\max(T_{k,n},\underline{T}_{k,n})}^{\min(T_{k,n},\bar{T}_{k,n})}(X_{k,t}-\mu_k)\big)\big(\frac{1}{T_{q,n}}\sum_{t'=\max(T_{q,n},\underline{T}_{q,n})}^{\min(T_{q,n},\bar{T}_{q,n})}(X_{q,t'}-\mu_q)\big)\Big]\Big|\\
&=\Big|\mathbb{E}\Big[\frac{1}{T_{k,n}}\frac{1}{T_{q,n}}Z\Big]\Big|\\
&=\Big|\sum_{t=\underline{T}_{k,n}}^{\bar{T}_{k,n}}\sum_{t'=\underline{T}_{q,n}}^{\bar{T}_{q,n}}Z\frac{1}{t}\frac{1}{t'}\mathbb{I}\{T_{k,n}=t,T_{q,n}=t'\}\Big|\\
&\leq \mathbb{E}[Z^2]\Big(\frac{1}{\underline{T}_{k,n}}\frac{1}{\underline{T}_{q,n}}-\frac{1}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}\Big).
\end{aligned}
$$

Note now that

$$
\begin{aligned}
\mathbb{E}[Z^2] &= \Big|\mathbb{E}\Big[\big(\sum_{t=\max(T_{k,n},\underline{T}_{k,n})}^{\min(T_{k,n},\bar{T}_{k,n})}(X_{k,t}-\mu_k)\big)\big(\sum_{t'=\max(T_{q,n},\underline{T}_{q,n})}^{\min(T_{q,n},\bar{T}_{q,n})}(X_{q,t'}-\mu_q)\big)\Big]\Big|\\
&\leq \sqrt{\mathbb{E}\Big[\big(\sum_{t=\max(T_{k,n},\underline{T}_{k,n})}^{\min(T_{k,n},\bar{T}_{k,n})}(X_{k,t}-\mu_k)\big)^2\Big]\mathbb{E}\Big[\big(\sum_{t'=\max(T_{q,n},\underline{T}_{q,n})}^{\min(T_{q,n},\bar{T}_{q,n})}(X_{q,t'}-\mu_q)\big)^2\Big]}\\
&\leq \sigma_k\sqrt{\bar{T}_{k,n}}\sigma_q\sqrt{\bar{T}_{q,n}}.
\end{aligned}
$$

From that, one gets

$$w_k w_q \Big| \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\tau\right\} \Big] \Big| \leq w_k \sigma_k \sqrt{\bar{T}_{k,n}} w_q \sigma_q \sqrt{\bar{T}_{q,n}} \Big(\frac{1}{\underline{T}_{k,n}} \frac{1}{\underline{T}_{q,n}} - \frac{1}{\bar{T}_{k,n}} \frac{1}{\bar{T}_{q,n}} \Big)$$

$$\leq 4A^2 \frac{\Sigma^2}{n^2} \frac{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}}{\bar{T}_{k,n}\bar{T}_{q,n}} \Big(\bar{T}_{k,n}\bar{T}_{q,n} - \underline{T}_{k,n}\underline{T}_{q,n} \Big) \tag{49}$$

$$\leq 4A^2 \frac{\Sigma^2}{n^2} \frac{1}{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}} \Big(\bar{T}_{k,n}\bar{T}_{q,n} - \underline{T}_{k,n}\underline{T}_{q,n} \Big). \tag{50}$$

where the second inequality comes from the fact that $\forall p, \underline{T}_{p,n} \geq T^*_{p,n} - A\lambda_p n^{2/3}$, which implies that $\frac{w_p \sigma_p}{\underline{T}_{p,n}} \leq \frac{\Sigma_w}{(n - A^{2/3})} \leq 2A\frac{\Sigma_w}{n}$.

Step 4: problem dependent upper bound We deduce from Equation 50 that

$$\begin{aligned}
& w_k w_q \Big| \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\tau\right\} \Big] \Big| \\
\leq & 4A^2 \frac{\Sigma_w^2}{n^2} \frac{1}{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}} \Big(\bar{T}_{k,n}\bar{T}_{q,n} - \underline{T}_{k,n}\underline{T}_{q,n} \Big) \\
\leq & 4A^2 \frac{\Sigma_w^2}{n^2} \frac{\Big(\big(\lambda_k n + D\sqrt{n}\big)\big(\lambda_q n + D\sqrt{n}\big) - \big(\lambda_k n - B\sqrt{n}\big)\big(\lambda_q n - B\sqrt{n}\big) \Big)}{\sqrt{\big(\lambda_k n + D\sqrt{n}\big)\big(\lambda_q n + D\sqrt{n}\big)}} \\
= & 4A^2 \frac{\Sigma_w^2}{n^2} \frac{\Big((D+B)(\lambda_p + \lambda_q) n\sqrt{n} + (D^2 - B^2) n \Big)}{\sqrt{\big(\lambda_k \lambda_q n^2 + (D+B)(\lambda_p + \lambda_q) n\sqrt{n} + D^2 n\big)}} \\
\leq & 4A^2 \frac{\Sigma_w^2}{n^2} \frac{(D + B + D^2) n\sqrt{n}}{n\sqrt{(\lambda_k \lambda_q)}} \\
\leq & 4A^2 \frac{(D + B + D^2)}{\sqrt{(\lambda_k \lambda_q)}} \frac{\Sigma_w^2}{n^{3/2}}.
\end{aligned}$$

Finally, we have

$$w_k w_q \Big| \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\tau\right\} \Big] \Big| \leq C_1 n^{-3/2}, \tag{51}$$

where $C_1 = 4A^2 \frac{(D+B+D^2)(\lambda_p + \lambda_q)}{\sqrt{(\lambda_k \lambda_q)}} \Sigma_w^2$.

Finally, using Equation 48, we have

$$\begin{aligned}
w_k w_q \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big) \Big] &= \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\xi\right\} \Big] + \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\xi^c\right\} \Big] \\
&\leq C_1 n^{-3/2} + C_\tau n^{-3/2}, \\
&\leq \big(C_1 + C_\tau\big) n^{-3/2},
\end{aligned}$$

where C_2 and C_τ depend only polynomially on $\log(n)$, on Σ_w, on K, on (c_1, c_2), and on $\frac{1}{\lambda_{\min}}$.

This concludes the proof for the problem dependent bound.

Step 4bis: problem independent upper bound From Equation 50, we deduce that

$$\begin{aligned}
&w_k w_q \Big| \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\tau\right\} \Big] \Big| \\
\leq & 16A^2 \frac{\Sigma_w^2}{n^2} \frac{1}{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}} \Big(\bar{T}_{k,n}\bar{T}_{q,n} - \underline{T}_{k,n}\underline{T}_{q,n} \Big) \\
\leq & 16A^2 \frac{\Sigma_w^2}{n^2} \frac{\Big(\big(\lambda_k n + Cn^{2/3}\big)\big(\lambda_q n + Cn^{2/3}\big) - \big(\lambda_k n - An^{2/3}\big)\big(\lambda_q n - An^{2/3}\big) \Big)}{\sqrt{\big(\lambda_k n + Cn^{2/3}\big)\big(\lambda_q n + Cn^{2/3}\big)}} \\
= & 16A^2 \frac{\Sigma_w^2}{n^2} \frac{\Big((A+C)(\lambda_p + \lambda_q) n n^{2/3} + (C^2 - A^2) n^{4/3} \Big)}{\sqrt{\big(\lambda_k\lambda_q n^2 + (A+C)(\lambda_p+\lambda_q) n n^{2/3} + C^2 n^{4/3}\big)}} \\
\leq & 16A^2 \frac{\Sigma_w^2}{n^2} \left[\frac{(A+C)(\lambda_p + \lambda_q) n n^{2/3}}{\sqrt{(A+C)(\lambda_p+\lambda_q) n n^{2/3}}} + \frac{(C^2 - A^2) n^{4/3}}{\sqrt{C^2 n^{4/3}}} \right] \\
\leq & 16A^2 \frac{\Sigma_w^2}{n^2} \left[\sqrt{(A+C)(\lambda_p + \lambda_q)} n^{5/6} + C n^{2/3} \right] \\
\leq & 16A^2 \left[\sqrt{(A+C)(\lambda_p + \lambda_q)} + C \right] \frac{\Sigma_w^2}{n^{7/6}}.
\end{aligned}$$

Finally, we have

$$w_k w_q \Big| \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\tau\right\} \Big] \Big| \leq C_2 n^{-7/6}, \tag{52}$$

where $C_2 = 16A^2 \left[\sqrt{(A+C)} + C \right] \Sigma_w^2$.

Finally, using Equation 48, we have

$$\begin{aligned}
w_k w_q \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big) \Big] &= \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\xi\right\} \Big] + \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\xi^c\right\} \Big] \\
&\leq C_2 n^{-7/6} + C_\tau n^{-3/2}, \\
&\leq \big(C_2 + C_\tau\big) n^{-7/6},
\end{aligned}$$

where C_2 and C_τ depend only polynomially on $\log(n)$, on Σ_w, on K and on (c_1, c_2).

This concludes the proof for the problem dependent bound.

■

References

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411:2712–2728, June 2010.

B. Arouna. Adaptative monte carlo method, a variance reduction technique. *Monte Carlo Methods and Applications*, 10(1):1–24, 2004.

J.Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *22nd annual conference on learning theory*, 2009.

J.Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

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

VV Buldygin and Y.V. Kozachenko. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32(6):483–489, 1980.

A. Carpentier, A. Lazaric, M. Ghavamzadeh, R. Munos, and P. Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *Algorithmic Learning Theory*, pages 189–203. Springer, 2011.

Alexandra Carpentier and Remi Munos. Finite time analysis of stratified sampling for monte carlo. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1278–1286. 2011.

Pierre Etoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodol. Comput. Appl. Probab.*, 12(3):335–360, September 2010.

Pierre Etoré, Gersende Fort, Benjamin Jourdain, and Éric Moulines. On adaptive stratification. *Ann. Oper. Res.*, 2011. to appear.

P. Glasserman. *Monte Carlo methods in financial engineering*. Springer Verlag, 2004. ISBN 0387004513.

P. Glasserman, P. Heidelberger, and P. Shahabuddin. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. *Mathematical Finance*, 9 (2):117–152, 1999.

V. Grover. Active learning and its application to heteroscedastic problems. *Department of Computing Science, Univ. of Alberta, MSc thesis*, 2009.

R. Kawai. Asymptotically optimal allocation of stratified sampling with adaptive variance reduction by strata. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 20(2):1–17, 2010. ISSN 1049-3301.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.

S.I. Resnick. *A probability path*. Birkhäuser, 1999.

R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo method*. Wiley-interscience, 2008. ISBN 0470177942.