# Finite-Time Analysis of Stratified Sampling for Monte Carlo

Alexandra Carpentier, Rémi Munos

# Finite Time Analysis of Stratified Sampling for Monte Carlo

**Anonymous Author(s)**
Affiliation
Address
email

## 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 regret analyses: a distribution-dependent bound $\widetilde{O}(n^{-3/2})$ that depends on a measure of the disparity of the strata, and a distribution-free bound $\widetilde{O}(n^{-4/3})$ that does not.

## 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 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 [1] 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 estimation of the true variances.

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 times their size[2]. There are however some open questions in this very good Master Thesis, notably on the asymptotic consistency of GAFS-WL and on the existence of a problem independent regret bound. We clarify this point in Section 4.

[**Jai change ici en parlant de GAFS-WL.**]

Our objective is similar, and we extend the analysis of this setting.

**Contributions:** In this paper, we introduce a new algorithm based on Upper-Confidence-Bounds (UCB) on the standard deviation. They are computed from the empirical standard deviation and a confidence interval derived from Bernstein's inequalities. We provide a finite-time analysis of its performance. The algorithm, called MC-UCB, samples the arms proportionally to an UCB[3] on the standard deviation times the size of the stratum. Our contributions are the following:

- We derive a *finite-time analysis* for the stratified sampling for Monte-Carlo setting by using an algorithm based on upper confidence bounds. By clarifying the notion of regret introduced in (Grover, 2009), we prove asymptotic consistency of our algorithm.
- We provide two regret analysis: (i) a distribution-dependent bound $\widetilde{O}(n^{-3/2})$[4] 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 $\widetilde{O}(n^{-4/3})$ 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

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

[2] This is explained in (Rubinstein and Kroese, 2008) and will be formulated precisely later.

[3] 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.

[4] The notation $\widetilde{O}(\cdot)$ corresponds to $O(\cdot)$ up to logarithmic factors.

the fact that the corresponding excess error rates differ enlightens the fact that a finite-time analysis is very relevant for this problem.

[**Jai change ici en parlant de consistence asymptotique.**]

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 introduces the MC-UCB algorithm and reports performance bounds. We then discuss in Section 4 about the parameters of the algorithm and its performances. In Section 5 we report numerical experiments that illustrate our method to the problem of pricing Asian options as introduced in (Glasserman et al., 1999). Finally, Section 6 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 $\nu_k$ with mean value $\mu_k$ and variance $\sigma_k^2$. At each round $t \geq 1$, an allocation strategy (or algorithm) $\mathcal{A}$ selects an arm $k_t$ and receives a sample drawn from $\nu_{k_t}$ 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 $\mu_k$ 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 $\mu$ as estimated by the weighted average $\hat{\mu}_n = \sum_{k=1}^K w_k \hat{\mu}_{k,n}$ satisfies:

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

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

$$L_n(\mathcal{A}) = \textstyle\sum_{k=1}^K w_k^2 \mathbb{E}\left[(\mu_k - \hat{\mu}_{k,n})^2\right]. \tag{1}$$

We give some properties of this performance in Appendix E.

[**Faut il expliquer et donner les deux propositions de l'Appendix?**]

The goal is to define an allocation strategy that minimizes the global loss defined in Equation 1. If the variance of the arms were known in advance, one could design an optimal static[5] 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 [6]

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

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 $k$,

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

[5] Static means that the number of pulls allocated to each arm does not depend on the received samples.

[6] 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.

times, and achieves a global performance

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

where $\Sigma_w = \sum_{i=1}^K w_i\sigma_i$. 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 $\lambda_{\min} = \min_{1\leq k\leq K} \lambda_k$. 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 3 a characterization of the hardness of a problem.

However, in the setting considered here, the $\sigma_k$ 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 $T_k^u = \frac{w_k}{\sum_{i=1}^K w_i} n$. Its performance 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 $\Sigma_w^2 \leq \Sigma_{w,2}$ with equality if and only if the $(\sigma_k)$ 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 $\sigma_k$ around their weighted mean $\Sigma_w$. In other words, it is the variance of the $(\sigma_k)_{1\leq k\leq K}$. As a result the gain of $\mathcal{A}^*$ compared to $\mathcal{A}^u$ grow with the disparity of the $\sigma_k$.

We would like to do better than the uniform strategy by considering an adaptive strategy $\mathcal{A}$ that would estimate the $\sigma_k$ 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.

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 *regret* of an adaptive algorithm $\mathcal{A}$ as the difference between the performance loss incurred by the algorithm and the optimal algorithm:

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

The *regret* indicates how much we loose in terms of expected quadratic estimation error by not knowing in advance the standard deviations $(\sigma_k)$. 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 neglectable compared to $1/n$.

## 3 Allocation based on Monte Carlo Upper Confidence Bound

### 3.1 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 $\delta$ which defines the *confidence level* of the bound. In Subsection 4.3, 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.

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 $\sigma_k$, for each arm $k$, and then pulls the one with largest $B_{k,t}$. The upper bounds on the standard deviations

**Input:** $c_1$, $c_2$, $\delta$. Let $b = 2\sqrt{2\log(2/\delta)}\sqrt{c_1\log(c_2/\delta)} + \frac{\sqrt{2c_1\delta(1+\log(c_2/\delta))}n^{1/2}}{(1-\delta)}$.
**Initialize:** Pull each arm twice.
**for** $t = 2K+1, \dots, n$ **do**
  Compute $B_{k,t} = \frac{w_k}{T_{k,t-1}}\left(\hat{\sigma}_{k,t-1} + b\sqrt{\frac{1}{T_{k,t-1}}}\right)$ 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 6.

are built by using Theorem 10 in (Maurer and Pontil, 2009)[7] and based on the empirical standard deviation $\hat{\sigma}_{k,t-1}$ :

$$\hat{\sigma}^2_{k,t-1} = \frac{1}{T_{k,t-1}-1}\sum_{i=1}^{T_{k,t-1}}(X_{k,i} - \hat{\mu}_{k,t-1})^2, \tag{6}$$

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 $1 \le k \le K$.

### 3.2 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 $1 \le k \le K$ and any $\epsilon > 0$,*

$$\mathbb{P}_{X\sim\nu_k}(|X - \mu_k| \ge \epsilon) \le c_2\exp(-\epsilon^2/c_1)\ . \tag{7}$$

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 4.

**A *distribution-dependent* result:** We now report the first bound on the regret of MC-UCB algorithm. The proof is reported in Appendix B (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 3).

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

$$R_n(\mathcal{A}_{MC-UCB}) \le \frac{\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}\left(112\Sigma_w\sqrt{c_1(c_2+2)}+6c_1(c_2+2)K\right)+\frac{19}{\lambda_{\min}^3 n^2}\left(K\Sigma_w^2+720c_1(c_2+1)\log(n)^2\right). \tag{8}$$

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.

**A *distribution-free* result:** Now we report our second regret bound that does not depend on $\lambda_{\min}$ but whose rate is poorer. The proof is given in Appendix C of the supplementary material and relies on other upper- and lower-bounds on $T_{k,t} - T^*_{k,t}$ detailed in Lemma 4.

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

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

$$R_n(\mathcal{A}_{MC-UCB}) \leq \frac{200\sqrt{c_1}(c_2+2)\Sigma_w K}{n^{4/3}}\log(n) + \frac{365}{n^{3/2}}\Big(129c_1(c_2+2)^2K^2\log(n)^2 + K\Sigma_w^2\Big). \tag{9}$$

This bound does not depend on $1/\lambda_{\min}$, not even in the $\widetilde{O}(\cdot)$ neglectable term, as detailed in Appendix C[8]. This is obtained at the price of the slightly worse rate $\widetilde{O}(n^{-4/3})$.

# 4 Discussion on the results

## 4.1 Asymptotic optimality

As explained in Section 2, our objective is to design a sampling strategy capable of estimating the mean value $\mu$ as accurately as possible given the strata. In order to assess our performance, we compare our strategy to the optimal static strategy $\mathcal{A}^*$ that requires the knowledge of the standard deviations in each stratum. Theorems 1 and 2 imply that the regret of MC-UCB is asymptotically neglectable compared to the optimal loss $L_n(\mathcal{A}^*) = \frac{\Sigma_w^2}{n}$. Thus, our algorithm 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 $\mu$ and the variance of $\hat{\mu}_n$ is asymptotically equal to the variance of the optimal allocation $\Sigma_w^2/n$. This comes from the fact that for algorithm MC-UCB, $\mathbb{E}(\hat{\mu}_n - \mu)^2 \leq L_n + \tilde{O}(n^{-3/2})$, where $\tilde{O}(\cdot)$ is $O(\cdot)$ up to $\lambda_{\min}^{-1}$ and $\log(n)$ factors. This is explained in Appendix E.2. Note that the asymptotic optimality of GAFS-WL is not provided in Grover (2009), although we believe it to hold also.

[**Jai change ici en rajoutant que on est pas sur que GAFS-WL soit optimal asymptotiquement et en expliquant pourquoi nous on lest.**]

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 better than the uniform strategy.

## 4.2 Distribution-free versus distribution-dependent

Theorem 1 provides a regret bound of order $\widetilde{O}(\lambda_{\min}^{-5/2}n^{-3/2})$, whereas Theorem 2 provides a bound in $\widetilde{O}(n^{-4/3})$ independently of $\lambda_{\min}$. Hence, for a given problem i.e., a given $\lambda_{\min}$, the distribution-free result of Theorem 2 is more informative than the distribution-dependent result of Theorem 1 in the *transitory regime*, that is to say when $n$ is small compared to $\lambda_{\min}^{-1}$. The distribution-dependent result of Theorem 1 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[9].

Although we do not have a lower bound on the regret yet, we believe that the rate $n^{-3/2}$ cannot be improved for general distributions. As explained in the proof in Appendix B, 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.* A natural question is whether there exists an algorithm with a regret of order $\widetilde{O}(n^{-3/2})$ without any dependence in $\lambda_{\min}^{-1}$. Although we do not have an answer to this question, we can say that our algorithm MC-UCB does not satisfy this property. In Appendix D.1 of the supplementary material, we give a simple example where $\lambda_{\min} = 0$ and for which the rate of MC-UCB cannot be better than $\widetilde{O}(n^{-4/3})$. This shows that our analysis of MC-UCB is tight.

[8]Note that the bound is not entirely distribution free since $\Sigma_w$ appears. But it can be proved using Assumption 1 that $\Sigma_w^2 \leq c_1c_2$.

[9]The distribution dependent bound is in $O(K\log n/\Delta)$, where $\Delta$ 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).

The problem dependent lower bound is similar to the one provided for GAFS-WL in (Grover, 2009). We however expect that GAFS-WL has for some problems a sub-optimal behavior: it is possible to find cases where $R_n(\mathcal{A}_{GAFS-WL}) \geq O(1/n)$, see Appendix D.2 for more details. 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.

[**Jai change ici en parlant dune lower bound problem dep pour GAFS-WL et en disant pquoi il est mieux que nous quand on a des bras a 0.**]

### 4.3 The parameters of the algorithm

Our algorithm takes three parameters as input, namely $c_1$, $c_2$ and $\delta$, but we only use a combination of them in the algorithm, with the introduction of $b = 2\sqrt{2\log(2/\delta)}\sqrt{c_1\log(c_2/\delta)} + \frac{\sqrt{2c_1\delta(1+\log(c_2/\delta))}n^{1/2}}{(1-\delta)}$. For practical use of the method, it is enough to tune the algorithm with a single parameter $b$. By the choice of the value assigned to $\delta$ in the two theorems, $b \approx c\log(n)$, 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, $b$ can be chosen as $b = 4\sqrt{\frac{5}{2}}c\sqrt{\log(n)}$ where $c$ is a true bound on the variables. This result is easy to deduce by simplifying Lemma 1 in Appendix A for the case of bounded variables.

## 5 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-Schole 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:

$$F(W) = \exp(-rT)\max\left[\frac{1}{d}\sum_{i=1}^{d} S_0\exp\left[(r-\frac{1}{2}s_0^2)\frac{iT}{d} + s_0\sqrt{T}W_i\right] - C, 0\right], \qquad (10)$$

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 (10), 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.

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

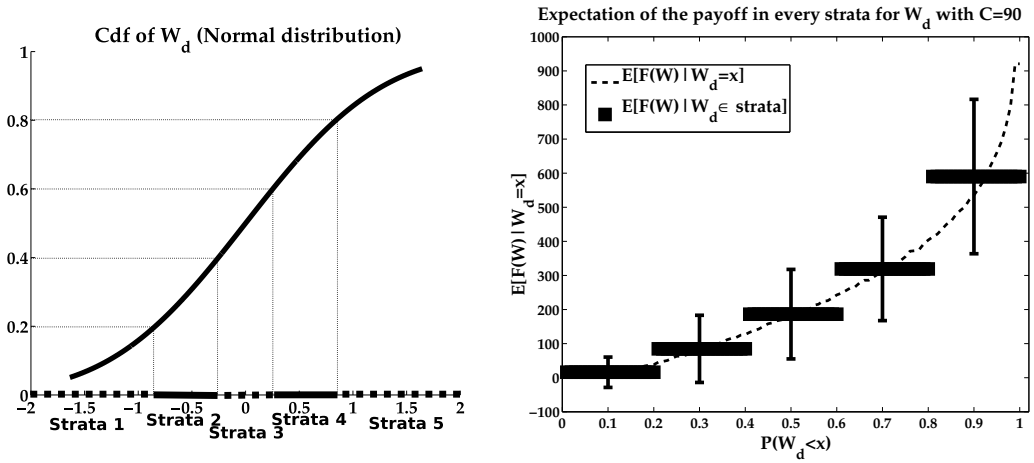


Figure 2: Left: Cdf of $W_d$ and the definition of the strata. Right: expectation and standard deviation of $F(W)$ conditioned on each stratum for a strike $C = 90$.

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[10]. 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 4.3, 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 5 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 SSAA and the uniform strategy and GAFS-WL strategy. 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).

[**Resultats pour GAFS-WL+Comparaison**]

[**Resultats pour GAFS-WL+Comparaison**]

In the left plot of Figure 3, we plot the rescaled regret $R_n n^{3/2}$, 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 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 regret

[10] 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.

| $C$ | $\frac{1}{\lambda_{\min}}$ | $\Sigma_{w,2}/\Sigma_w^2$ | 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 $C$.

more accurately. The curve grows first and then stabilizes. This could correspond to the two regimes discussed previously.

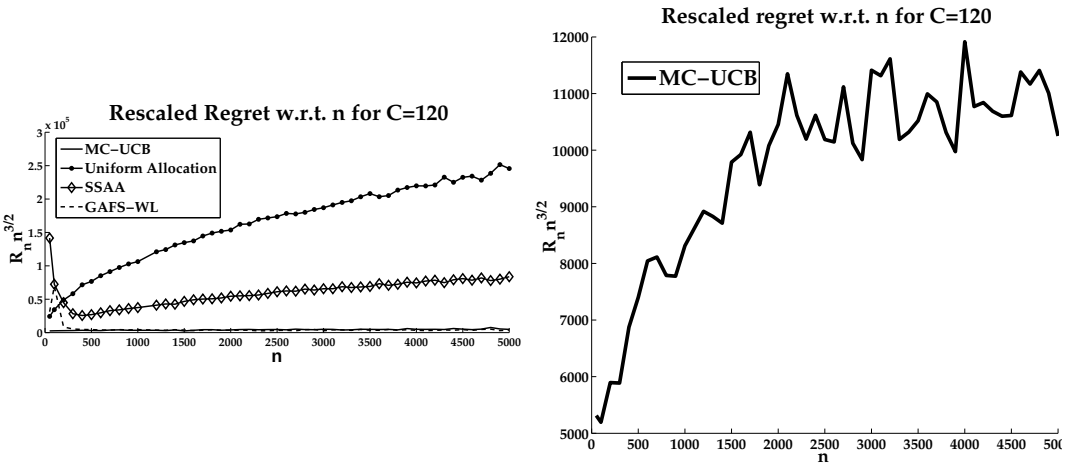


Figure 3: Left: Rescaled regret ($R_n n^{3/2}$) of the Uniform, SSAA, and MC-UCB strategies. Right: zoom on the rescaled regret for MC-UCB that illustrates the two regimes.

## 6 Conclusions

We provide a finite-time analysis for stratified sampling for Monte-Carlo in the case of fixed strata. 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 reported two bounds: (i) a distribution dependent bound $\widetilde{O}(n^{-3/2}\lambda_{\min}^{-5/2})$ 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 in $\widetilde{O}(n^{-4/3})$ which is of interest when $n$ is small compared to $\lambda_{\min}^{-1}$ (*transitory regime*).

Possible directions for future work include: (i) making the MC-UCB algorithm anytime (i.e. not requiring the knowledge of $n$), (ii) investigating whether their exists an algorithm with $\widetilde{O}(n^{-3/2})$ regret without dependency on $\lambda_{\min}^{-1}$, and (iii) deriving distribution-dependent and distribution-free lower-bounds for this problem.

## 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.

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.

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

## A Main tools

### A.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 1** *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 6 for $t$ samples. The event $\xi$ 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

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

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 $\sigma^2$ and mean $\mu$ 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 $\xi_1$, 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 1 (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 1** *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 $\nu_k$, 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 6. Then, on the event $\xi$, we have:*

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

### A.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 1** *Let $\{\mathcal{F}_t\}$ be a filtration and $X_t$ a $\mathcal{F}_t$-adapted sequence of i.i.d. random variables with variance $\sigma^2$. Assume that $\mathcal{F}_t$ and the $\sigma$-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 the regret outside of $\xi$.** The next lemma provides a bound for the loss whenever the event $\xi$ does not hold.

**Lemma 2** *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 $\epsilon = c_1 \log(c_2/2nK\delta)$[11], we obtain

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

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

We thus know that

$$\begin{aligned}
&\max_{\Omega/\mathbb{P}(\Omega)=2nK\delta} \mathbb{E}\big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{\Omega\right\}\big] \\
&\leq \int_{c_1\log(c_2/2nK\delta)}^{\infty} c_2\exp(-\epsilon/c_1)d\epsilon + c_1\log(c_2/2nK\delta)\mathbb{P}\Big(\Omega\Big) \\
&= 2c_1nK\delta(1+\log(c_2/2nK\delta))\ .
\end{aligned}$$

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

$$\mathbb{E}\big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq \max_{\Omega/\mathbb{P}(\Omega)=2nK\delta} \mathbb{E}\big[|X_{k,t}-\mu_k|^2\mathbb{I}\left\{\Omega\right\}\big] \leq 2c_1nK\delta(1+\log(c_2/2nK\delta))\ .$$

The claim follows from the fact that $\mathbb{E}\big[|\hat{\mu}_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq \sum_{t=1}^{n}\mathbb{E}\big[|X_{k,n}-\mu_k|^2\mathbb{I}\left\{\xi^C\right\}\big] \leq 2c_1n^2K\delta(1+\log(c_2/2nK\delta))$. □

### A.3 Technical inequalities

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

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

[**On multiplie juste par $\sqrt{2}$ et on vire le terme en $\hat{\mu}$ par rapport a avant.**]

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

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

**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

$$\mathbb{E}\big[|X_{k,t}-\mu_k|^2\big] \leq \int_0^{\infty} c_2\exp(-\epsilon/c_1)d\epsilon = c_2c_1$$

We then have $\Sigma_w\leq\sqrt{c_2c_1}$.

If $\delta = n^{-7/2}$, we obtain by using the lower bound on $a$ that

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

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 2 when $\delta = n^{-7/2}$ and when choosing $c_2$ such that $c_2 \geq 2enK\delta = 2Ken^{-5/2}$

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

# B Proof of Theorem 1

In this section, we first provide the proof for the following Lemma and then use the result to prove Theorem 1.

**Lemma 3** *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 $\xi$ (and thus with probability at least $1 - 2nK\delta$), for any arm $1 \leq p \leq K$:*

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

*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 $\sigma_k$, which is of order $1/\sqrt{n}$. Note that Equation 17 also shows the inverse dependency on the smallest proportion $\lambda_{\min}$.

*Proof:* [Lemma 3] 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 1, we obtain the following upper and lower bounds for $B_{q,t+1}$ on $\xi$:

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

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 $\xi$).

**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 $\xi$ 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. □

We are now ready to prove Theorem 1.

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

$$R_n(\mathcal{A}_{MC-UCB}) \leq \frac{\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}\Big(112\Sigma_w\sqrt{c_1(c_2+2)} + 6c_1(c_2+2)K\Big) + \frac{19}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 720c_1(c_2+1)\log(n)^2\Big) .$$

*Proof:* [Theorem 1] 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. Regret bound.** By definition, the loss of the algorithm writes

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

Using the definition of $\hat{\mu}_{k,n}$ and Proposition 1 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 $\xi$.

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 $\xi$), 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 2, 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

$$\begin{aligned}
L_n &\leq \frac{\Sigma_w^2}{n} + 24a\Sigma_w\frac{\sqrt{\log(2/\delta)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{16}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 18a^2\log(2/\delta)\Big) + 2c_1n^2K\delta(1+\log(c_2/2nK\delta)) \\
&\leq \frac{\Sigma_w^2}{n} + 54a\Sigma_w\frac{\sqrt{\log(n)}}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{16}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 90a^2\log(n)\Big) + 6c_1K(c_2+1)\log(n)n^{-3/2} \\
&\leq \frac{\Sigma_w^2}{n} + \frac{\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}\Big(112\Sigma_w\sqrt{c_1(c_2+2)} + 6c_1(c_2+2)K\Big) + \frac{19}{\lambda_{\min}^3 n^2}\Big(K\Sigma_w^2 + 720c_1(c_2+1)\log(n)^2\Big)\,.
\end{aligned}$$

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

The Theorem follows by expressing the regret.

□

# C Proof of Theorem 2

Again, we first state and prove the following Lemma and then use this result to prove Theorem 2.

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

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

*and*

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

*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 3, 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:*

**Step 1. Lower bound of order $\widetilde{O}(n^{2/3})$.** Let $k$ be the index of an arm such that $T_{k,n} \geq \frac{n}{K}$ (this implies $T_{k,n} \geq 3$ as $n \geq 4K$, and arm $k$ is thus pulled after the initialization) and let $t+1 \leq n$ be the last time at which it was pulled [12], i.e., $T_{k,t} = T_{k,n} - 1$ and $T_{k,t+1} = T_{k,n}$. From Equation 15 and the fact that $T_{k,n} \geq \frac{n}{K}$, we obtain on $\xi$

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

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

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

[12] Note that such an arm always exists for any possible allocation strategy given the constraint $n = \sum_q T_{q,n}$.

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

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

Combining Equations 30–32, we obtain on $\xi$

$$2aw_q \frac{\sqrt{\log(2/\delta)}}{T_{q,n}^{3/2}} \leq \frac{K\Big(\Sigma_w + 4a\sqrt{\log(2/\delta)}\Big)}{n}.$$

Finally, this implies on $\xi$ that for any $q$,

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

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

$$c(\delta) = \Big(\frac{2a\sqrt{\log(2/\delta)}}{\Sigma_w + 4a\sqrt{\log(2/\delta)}} \frac{1}{K}\Big)^{2/3},$$

thus the lower bound on $T_{q,n}$ on $\xi$ 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 3. 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 1 it follows that, on $\xi$

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

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 $\xi$

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

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

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

Combining Equations 35–37, we obtain on $\xi$

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

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 $\xi$

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

This implies

$$\frac{w_p\sigma_p}{T_{p,n}}(n-3K) \leq \sum_{q=1}^{K} w_q\left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}}\right). \tag{38}$$

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

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

on $\xi$, since $\sum_q w_q^{2/3} \leq K^{2/3}$ 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 $\xi$ the following bound

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

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

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

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 $\xi$ follows by using $T_{p,n} = n - \sum_{q\neq p} T_{q,n}$ and the previous lower bound, that is

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

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

We are now ready to prove Theorem 2.

**Theorem 2** *Under Assumption 1 and by ensuring that $c_2 \geq 2Kn^{-5/2}$, the regret of MC-UCB run with parameter $\delta = n^{-7/2}$ with $n \geq 4K$ is bounded as*

$$R_n(\mathcal{A}_{MC-UCB}) \leq \frac{200\sqrt{c_1}(c_2+2)\Sigma_w K}{n^{4/3}}\log(n) + \frac{365}{n^{3/2}}\Big(129c_1(c_2+2)^2K^2\log(n)^2 + K\Sigma_w^2\Big).$$

*Proof:* [Theorem 2]

We decompose the loss on $\xi$ and its complement:

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

Using the definition of $\hat{\mu}_{k,n}$ and Proposition 1 we bound the first term as

$$\sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\}\Big] \leq \sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2 \mathbb{E}[T_{k,n}]}{\underline{T}_{k,n}^2}\,, \tag{40}$$

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

Note also that as $\sum_k T_{k,n} = n$, we also have $\sum_k \mathbb{E}[T_{k,n}] = n$. Using Equation 40 and Equation 39 which provides an upper bound on $\xi$ on $\frac{w_k \sigma_k}{T_{k,n}}$ (and thus a lower bound on $\xi$ on $T_{k,n}$), we deduce

$$\sum_{k=1}^{K} w_k^2 \mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\left\{\xi\right\}\Big] \leq \sum_{k=1}^{K} \Big(\frac{\Sigma_w}{n} + 24aK^{2/3}\sqrt{\frac{\log(2/\delta)}{c(\delta)}} n^{-4/3} + \frac{12K\Sigma_w}{n^2}\Big)^2 \mathbb{E}[T_{k,n}]. \tag{41}$$

Using the fact that $\sum_k \mathbb{E}[T_{k,n}] = n$, Equation 41 may be rewritten as

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

From Lemma 2, 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

$$\begin{aligned}
L_n &\leq \frac{\Sigma_w^2}{n} + \frac{48\Sigma_w aK^{2/3}}{n^{4/3}}\sqrt{\frac{\log(2/\delta)}{c(\delta)}} + \frac{300}{n^2}\Big(4a^2K^{4/3}\frac{\log(2/\delta)}{c(\delta)} + K\Sigma_w^2\Big) + 2c_1 n^2 K\delta(1 + \log(c_2/2nK\delta)) \\
&\leq \frac{\Sigma_w^2}{n} + \frac{96\Sigma_w aK}{n^{4/3}}\sqrt{\log(n)}\big(\sqrt{c_2}+1\big)^{1/3} + \frac{300}{n^2}\Big(16a^2K^2\log(n)\big(\sqrt{c_2}+1\big)^{2/3} + K\Sigma_w^2\Big) \\
&+ 6c_1K(c_2+1)\log(n)n^{-3/2} \\
&\leq \frac{\Sigma_w^2}{n} + \frac{200\sqrt{c_1(c_2+2)}\Sigma_w K}{n^{4/3}}\log(n)\big(\sqrt{c_2}+1\big)^{1/3} \\
&+ \frac{365}{n^{3/2}}\Big(16a^2K^2\log(n)\big(\sqrt{c_2}+1\big)^{2/3} + K\Sigma_w^2 + c_1(c_2+2)K\log(n)\Big) \\
&\leq \frac{\Sigma_w^2}{n} + \frac{200\sqrt{c_1}(c_2+2)\Sigma_w K}{n^{4/3}}\log(n) + \frac{365}{n^{3/2}}\Big(129c_1(c_2+2)^2K^2\log(n)^2 + K\Sigma_w^2\Big)\,.
\end{aligned}$$

where we use $a \leq 2\sqrt{2c_1(c_2+2)\log(n)}$, $c(\delta) = \Big(\frac{1}{2K}\Big)^{2/3}\Big(\frac{1}{\sqrt{c_2}+1}\Big)^{2/3}$ 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 A.3.

□

# D Comments on problem independent bounds for MC-UCB and GAFS-WL

## D.1 Note on the problem independent bound for MC-UCB

An interesting question is whether it is possible to obtain a regret bound of order $n^{-3/2}$ without the dependency on $\lambda_{\min}^{-1}$. We provide a simple example that demonstrates that MC-UCB does not possess this property.

Consider a problem with $K = 2$ arms with $\sigma_1 = 1$ and $\sigma_2 = 0$ and $w_1 = w_2 = 0.5$. The optimal allocation strategy for this problem is $T^*_{1,n} = n - 1$, $T^*_{2,n} = 1$ (we need only one sample of the second arm to estimate its mean). Since $\lambda_{\min} = 0$ the bound in Theorem 1 is meaningless (although MC-UCB is still able to minimize the regret as demonstrated by Theorem 2). Indeed, the definition of the upper-confidence bound on the standard deviation forces the algorithm to pull each arm at least $\widetilde{O}(n^{2/3})$ times, including those arms with zero variance. Hence in this example, arm 2 will be pulled $\widetilde{O}(n^{2/3})$ times, which results in under-pulling arm 1 by the same amount, and thus, in worsening its estimation. It can be shown that the resulting regret is $\widetilde{O}(n^{-4/3})$, which still decreases to zero faster than $1/n$ but with a poorer rate. A sketch of the proof of this argument is as follows. Using the definition of $B_{k,t}$ and Equation 20 (see Appendix B), since $\hat{\sigma}_2 = 0$, we have that at any time $t + 1 > 2$

$$B_{1,t+1} \leq \frac{1}{2T_{1,t}}\Big(1 + b\Big) \qquad \text{and} \qquad B_{2,t+1} = \frac{1}{2T_{2,t}}\Big(b\sqrt{\frac{1}{T_{2,t}}}\Big). \tag{42}$$

Let $t+1 \leq n$ be the last time that arm 1 was pulled, i.e., $T_{1,t} = T_{1,n}-1$ and $B_{1,t+1} \geq B_{2,t+1}$. From Equation 42, we have

$$B_{2,t+1} = \frac{b}{2T_{2,t}^{3/2}} \leq B_{1,t+1} \leq \frac{1}{2T_{1,n} - 1}\Big(1 + b\Big). \tag{43}$$

Now consider the two possible cases: **1)** $T_{1,n} \leq n/2$, in which case obviously $T_{2,n} \geq n/2$ and **2)** $T_{1,n} > n/2$, in this case Equation 43 implies that $T_{2,n} \geq T_{2,t} = \widetilde{O}\big(n^{2/3}\big)$. Thus in both cases, we may write $T_{2,n} = \widetilde{O}\big(n^{2/3}\big)$, which indicates that arm 2 (resp. arm 1) is over-sampled (resp. under-sampled) by a number of pulls of order $\widetilde{O}(n^{2/3})$. By following the same arguments as in the proof of Theorem 2, we deduce that the regret in this case is of order $\widetilde{O}(n^{-4/3})$. Note that this poorer rate is the result of over-sampling the arm with the smaller variance (and as a consequence under-sampling at least one arm with a larger variance).

Thus in the case of an arm with 0 standard deviation, the regret of MC-UCB is at least $\widetilde{O}(n^{-4/3})$.

## D.2 Note for a 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 $\epsilon$ 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

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

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

$$\begin{aligned}L_n(\mathcal{A}_{GAFS-WL}) &\geq \frac{1}{4n} + \mathbb{P}(\hat{\sigma}_{1,a\sqrt{n}}=0)\mathbb{P}(\hat{\sigma}_{2,a\sqrt{n}}>0)\Big(n^{-1/2-\epsilon}\Big)^2 \\ &\geq \frac{1}{4n} + (1-2\exp(-a\sqrt{n}/4))(1-an^{-\epsilon})^2\Big(n^{-1-2\epsilon}\Big) \\ &\geq \frac{1}{4n} + (1-\frac{8}{a\sqrt{n}})(1-2\frac{a}{n^{\epsilon}})\frac{1}{n^{1+2\epsilon}} \\ &\geq \frac{1}{4n} + \frac{1}{n^{1+2\epsilon}} - \frac{8}{an^{3/2+2\epsilon}} - \frac{2a}{n^{1+3\epsilon}} \\ &\geq \frac{1}{4n} + \frac{1}{n^{1+2\epsilon}} - \frac{10\max(a,1/a)}{n^{1+3\epsilon}},\end{aligned}$$

where the last line is obtained using the fact that $\epsilon < 1/6$.

The loss thus verifies

$$\begin{aligned}R_n(\mathcal{A}_{GAFS-WL}) &\geq \frac{1}{n^{1+2\epsilon}} - \frac{10\max(a,1/a)}{n^{1+3\epsilon}} - \frac{1}{n^{5/4+\epsilon/2}} \\ &\geq \frac{1}{n^{1+2\epsilon}} - \frac{11\max(a,1/a)}{n^{1+3\epsilon}},\end{aligned}$$

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

$$R_n(\mathcal{A}_{GAFS-WL}) \geq \frac{1}{2n^{1+2\epsilon}},$$

with $\epsilon$ arbitrarily close to 0.

# E Some properties of the regret

## E.1 A property of the regret in the special case of symmetric distributions

**Proposition 2** *f the distribution $\nu_k$ of the arms are symmetric around $\mu_k$ respectively, then for the algorithm MC-UCB*

$$L_n(\mathcal{A}_{MC-UCB}) = \mathbb{E}[(\hat{\mu}_n - \mu)^2],$$

*where the expectation of the right hand term is on the samples when running MC-UCB.*

*Proof:*

**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:

$$\begin{aligned}
&\mathbb{E}[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)|T_{k,n} = T_1, T_{q,n} = T_2] \\
&= \mathbb{E}\Big[\big(\frac{1}{T_1}\sum_{u=1}^{T_1} X_{k,u} - \mu_k\big)\big(\frac{1}{T_2}\sum_{u=1}^{T_2} X_{q,u} - \mu_q\big)|T_{k,n} = T_1, T_{q,n} = T_2\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\big(\frac{1}{T_1}\sum_{u=1}^{T_1} X_{k,u} - \mu_k\big)\big(\frac{1}{T_2}\sum_{u=1}^{T_2} X_{q,u} - \mu_q\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big] \\
&\times \mathbb{P}\big(\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}|T_{k,n} = T_1, T_{q,n} = T_2\big)|T_{k,n} = T_1, T_{q,n} = T_2\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\big(\frac{1}{T_1}\sum_{u=1}^{T_1} X_{k,u} - \mu_k\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big]\mathbb{P}\big(\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}|T_{k,n} = T_1, T_{q,n} = T_2\big)|T_{k,n} = T_1, T_{q,n} = T_2\Big] \\
&\times \mathbb{E}\Big[\mathbb{E}\big[\big(\frac{1}{T_2}\sum_{u=1}^{T_2} X_{q,u} - \mu_q\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big]\mathbb{P}\big(\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}|T_{k,n} = T_1, T_{q,n} = T_2\big)|T_{k,n} = T_1, T_{q,n} = T_2\Big],
\end{aligned} \tag{44}$$

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 $\nu_k$ 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 $\nu_k$ is symmetric around $\mu_k$.

This leads to

$$\mathbb{E}\big[\big(\frac{1}{T}\sum_{u=1}^{T} X_{k,u} - \mu_k\big)|\{\hat{\sigma}_{p,t'}\}_{p,t'\leq n}\big] = 0. \tag{45}$$

**Step 4: The cross products $\mathbb{E}[(\hat{\mu}_{k,n}-\mu_k)(\hat{\mu}_{q,n}-\mu_q)]$ are null.** We combine Equations 44 and 45 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

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

□

### E.2 A distribution dependent bound in the general case

**Proposition 3** *For algorithm MC-UCB when run with the parameters as in Theorem 1 and 2 , we have*

$$\mathbb{E}[(\hat{\mu}_n-\mu)^2]\leq L_n+\tilde{O}(n^{-3/2}),$$

*where $\tilde{O}(\cdot)$ is a $O(\cdot)$ up to $\lambda_{\min}^{-1}$ and $\log(n)$ factors.*

*Proof:*

**Step 0: A useful Lemma.**

**Lemma 5** *Let $(X,Y)$ be a couple of random variables such that $\mathbb{E}(XY)=0$. 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*

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

*Proof:*

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

$$XY\sum_{u=1}^{p}a_u\mathbb{I}\left\{(X,Y)\in\Omega_u\right\}\leq XYa_1\mathbb{I}\left\{XY\geq 0\right\}+XYa_p\mathbb{I}\left\{XY<0\right\}.$$

This implies

$$\begin{aligned}
\mathbb{E}\Big[XY\sum_{u=1}^{p}a_u\mathbb{I}\left\{(X,Y)\in\Omega_u\right\}\Big] &\leq \mathbb{E}\Big[XYa_1\mathbb{I}\left\{XY\geq 0\right\}+XYa_p\mathbb{I}\left\{XY<0\right\}\Big]\\
&\leq \mathbb{E}\Big[(a_1-a_p)XY\mathbb{I}\left\{XY\geq 0\right\}+a_pXY(\mathbb{I}\left\{XY<0\right\}+\mathbb{I}\left\{XY\geq 0\right\})\Big]\\
&\leq (a_1-a_p)\mathbb{E}\Big[XY\mathbb{I}\left\{XY\geq 0\right\}\Big]\\
&\leq (a_1-a_p)\sqrt{\mathbb{E}\Big[X^2\mathbb{I}\left\{XY\geq 0\right\}\Big]}\sqrt{\mathbb{E}\Big[Y^2\mathbb{I}\left\{XY\geq 0\right\}\Big]}\\
&\leq (a_1-a_p)\sqrt{\mathbb{E}\Big[X^2\Big]}\sqrt{\mathbb{E}\Big[Y^2\Big]},
\end{aligned}$$

by Cauchy-Schwartz.

By remarking that

$$XY\sum_{u=1}^{p}a_u\mathbb{I}\left\{(X,Y)\in\Omega_u\right\}\geq XYa_1\mathbb{I}\left\{XY\leq 0\right\}+XYa_p\mathbb{I}\left\{XY>0\right\},$$

we show in the same way that

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

Those two inequalities lead to the desired result.

□

**Step 1: Expressing the cross-products $\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)(\hat{\mu}_{q,n}-\mu_q)\Big]$.** Let $k$ and $q$ with $k\neq q$ be two arms. Let us call $X_{p,u}$ the $u$-th sample from arm $p$. We have if $n\geq 2K$

$$\begin{aligned}
&\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)\\
&=\sum_{T_1=2}^{n}\sum_{T_2=2}^{n}\Big[\big(\frac{1}{T_1}\sum_{u=1}^{T_1}X_{k,u}-\mu_k\big)\Big]\Big[\big(\frac{1}{T_2}\sum_{u=1}^{T_2}X_{q,u}-\mu_q\big)\Big]\mathbb{I}\left\{T_{k,n}=T_1,T_{q,n}=T_2\right\}.
\end{aligned}$$

**Step 2: Some additional properties of $\xi$.** *Number of pulls:* We remind that, on $\xi$ we have for all $p$,

$$T_{p,n}\geq \underline{T}_{p,n}=\max\Big(T^*_{p,n}-\min(A\lambda_p n^{2/3},Bn^{1/2}),En^{2/3}\Big),$$

and

$$T_{p,n}\leq \bar{T}_{p,n}=T^*_{p,n}+\min(Cn^{2/3},Dn^{1/2}),$$

where $A$, $B$, $C$ and $D$ are as in Theorem 1 and 2 and $E$ is as in the proof of Theorem 2. Note that $B$ and $D$ depend of $\lambda_{\min}$.

*Cross-products:*

We have also

$$\begin{aligned}
&|\mathbb{E}[(X_{q,v}-\mu_q)(X_{k,u}-\mu_k)\mathbb{I}\{\xi^c\}]| \\
&\leq \sqrt{\mathbb{E}[(X_{q,v}-\mu_q)^2\mathbb{I}\{\xi^c\}]}\sqrt{\mathbb{E}[(X_{k,u}-\mu_k)^2\mathbb{I}\{\xi^c\}]} \\
&\leq 2c_1nK\delta(1+\log(c_2/2nK\delta)) \leq \frac{5c_1(c_2+2)K\log(n)}{n^{3/2}},
\end{aligned}$$

as in Appendix A. And thus, as $\mathbb{E}[(X_{q,v}-\mu_q)(X_{k,u}-\mu_k)\mathbb{I}\{\xi\}] + \mathbb{E}[(X_{q,v}-\mu_q)(X_{k,u}-\mu_k)\mathbb{I}\{\xi^c\}] = 0$ because of $X_{q,v}$ is independent of $X_{k,u}$, we have

$$|\mathbb{E}[(X_{q,v}-\mu_q)(X_{k,u}-\mu_k)\mathbb{I}\{\xi\}]| \leq \frac{5c_1(c_2+2)K\log(n)}{n^{3/2}},$$

Let us note $r = \frac{5c_1(c_2+2)K\log(n)}{n^{3/2}}$.

The same way as before, for any $(T_1, T_2)$,

$$|\mathbb{E}[(\frac{1}{T_1}\sum_{u=1}^{T_1} X_{k,u}-\mu_k)(\frac{1}{T_2}\sum_{v=1}^{T_2} X_{q,v}-\mu_q)\mathbb{I}\{\xi\}]| \leq r,$$

and

$$|\mathbb{E}[(X_{k,u}-\mu_k)(\frac{1}{T_2}\sum_{v=1}^{T_2} X_{q,v}-\mu_q)\mathbb{I}\{\xi\}]| \leq r.$$

*Cross-products on specific events:*

Let us pose $t' \geq 1$. Let us fix $v$ such that $v \geq T_2$. Note that by the structure of the algorithm, $X_{q,v}$ is independent of all the samples of all arms on the event $\{T_{q,n} = T_2\}$. Thus, $\mathbb{E}\Big[\big(\frac{1}{T}\sum_{u=1}^{T} X_{k,u}-\mu_k\big)\big(X_{q,v}-\mu_q\big)\mathbb{I}\{T_{k,n}=T_1, T_{q,n}=T_2\}\Big] = 0$ for any $T_1$, any $T$ and any $k \neq q$. We conclude that

$$\Big|\mathbb{E}\Big[\big(\frac{1}{T}\sum_{u=1}^{T} X_{k,u}-\mu_k\big)\big(X_{q,v}-\mu_q\big)\mathbb{I}\{T_{k,n}=T_1, T_{q,n}=T_2\}\,\mathbb{I}\{\xi\}\Big]\Big| \leq r, \tag{46}$$

as Equation $\Big|\mathbb{E}\Big[\big(\frac{1}{T}\sum_{u=1}^{T} X_{k,u}-\mu_k\big)\big(X_{q,v}-\mu_q\big)\mathbb{I}\{T_{k,n}=T_1, T_{q,n}=T_2\}\,\mathbb{I}\{\xi^c\}\Big]\Big| \leq r$ holds.

In the same way, we obtain when $v \geq T_2$, we have for any $u$, any $k \neq q$

$$\Big|\mathbb{E}\Big[\big(X_{k,u}-\mu_k\big)\big(X_{q,v}-\mu_q\big)\mathbb{I}\{T_{k,n}=T_1, T_{q,n}=T_2\}\,\mathbb{I}\{\xi\}\Big]\Big| \leq r \tag{47}$$

**Step 3: Expressing** $(\hat{\mu}_{k,n}-\mu_k)(\hat{\mu}_{q,n}-\mu_q)\mathbb{I}\{\xi\}$**.** Now note that on $\xi$, for any arm $p$, $\underline{T}_{p,n} \leq T_{p,n} \leq \bar{T}_{p,n}$.

Let us consider the following sub-term of the previous equation:

$$
\begin{aligned}
&\big(\hat{\mu}_{k,n}-\mu_k\big)\big(\hat{\mu}_{q,n}-\mu_q\big)\mathbb{I}\{\xi\}\\
&=\sum_{T_1=\underline{T}_{k,n}}^{\bar{T}_{k,n}}\sum_{T_2=\underline{T}_{q,n}}^{\bar{T}_{q,n}}\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)\mathbb{I}\{T_{k,n}=T_1,T_{q,n}=T_2\}\,\mathbb{I}\{\xi\}\\
&=\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\big(\frac{1}{\underline{T}_{k,n}+t}\sum_{u=1}^{\underline{T}_{k,n}+t}(X_{k,u}-\mu_k)\big)\\
&\times\big(\frac{1}{\underline{T}_{q,n}+t'}\sum_{v=1}^{\underline{T}_{q,n}+t'}(X_{q,v}-\mu_q)\big)\mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\big\}\,\mathbb{I}\{\xi\}\\
&=\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\big(\frac{1}{\underline{T}_{k,n}+t}\sum_{u=1}^{\underline{T}_{k,n}}X_{k,u}-\mu_k\big)+\frac{1}{\underline{T}_{k,n}+t}\sum_{u=\underline{T}_{k,n}+1}^{\underline{T}_{k,n}+t}(X_{k,u}-\mu_k)\big)\Big]\\
&\times\Big[\big(\frac{1}{\underline{T}_{q,n}+t'}\sum_{v=1}^{\underline{T}_{q,n}}X_{q,v}-\mu_q\big)+\frac{1}{\underline{T}_{q,n}+t'}\sum_{v=\underline{T}_{q,n}+1}^{\underline{T}_{q,n}+t'}(X_{q,v}-\mu_q)\big)\Big]\\
&\times\mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\big\}\,\mathbb{I}\{\xi\}\\
&=\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\big(m_k-\mu_k\big)+\frac{1}{\underline{T}_{k,n}+t}\sum_{u=\underline{T}_{k,n}+1}^{\underline{T}_{k,n}+t}(X_{k,u}-\mu_k)\big)\Big]\\
&\times\Big[\frac{\underline{T}_{q,n}}{\underline{T}_{q,n}+t'}\big(m_q-\mu_q\big)+\frac{1}{\underline{T}_{q,n}+t'}\sum_{v=\underline{T}_{q,n}+1}^{\underline{T}_{q,n}+t'}(X_{q,v}-\mu_q)\big)\Big]\\
&\times\mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\big\}\,\mathbb{I}\{\xi\}\\
&=\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\big(m_k-\mu_k\big)\frac{\underline{T}_{q,n}}{\underline{T}_{q,n}+t'}\big(m_q-\mu_q\big)\Big]\\
&\times\mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\big\}\,\mathbb{I}\{\xi\} \qquad (48)\\
&+\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\big(m_k-\mu_k\big)\big(\frac{1}{\underline{T}_{q,n}+t'}\sum_{v=\underline{T}_{q,n}+1}^{\underline{T}_{q,n}+t'}(X_{q,v}-\mu_q)\big)\Big]\\
&\times\mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\big\}\,\mathbb{I}\{\xi\} \qquad (49)\\
&+\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{q,n}}{\underline{T}_{q,n}+t'}\big(m_q-\mu_q\big)\big(\frac{1}{\underline{T}_{k,n}+t}\sum_{u=\underline{T}_{k,n}+1}^{\underline{T}_{k,n}+t}(X_{k,u}-\mu_k)\big)\Big]\\
&\times\mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\big\}\,\mathbb{I}\{\xi\} \qquad (50)\\
&+\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\big(\frac{1}{\underline{T}_{q,n}+t'}\sum_{v=\underline{T}_{q,n}+1}^{\underline{T}_{q,n}+t'}(X_{q,v}-\mu_q)\big)\big(\frac{1}{\underline{T}_{k,n}+t}\sum_{u=\underline{T}_{k,n}+1}^{\underline{T}_{k,n}+t}(X_{k,u}-\mu_k)\big)\Big]\\
&\times\mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\big\}\,\mathbb{I}\{\xi\}\,. \qquad (51)
\end{aligned}
$$

where $m_p=\frac{1}{\underline{T}_{p,n}}\sum_{u=1}^{\underline{T}_{p,n}}X_{p,u}$.

**Step 4.1: Bounding the first term (Equation 48).** Let us pose $c=\mathbb{E}\Big[(m_k-\mu_k)(m_q-\mu_q)\mathbb{I}\{\xi\}\Big]$.

We want to bound (Equation 48)

$$\left|\mathbb{E}\Big[\big(m_q-\mu_q\big)\big(m_k-\mu_k\big)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{\underline{T}_{q,n}}{\underline{T}_{q,n}+t'}\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\right|$$

By Lemma 6 where the partition of $\xi$ is $\Big\{\xi\cap\{T_{k,n}=T_1,T_{q,n}=T_2\}\Big\}_{T_1=\underline{T}_{k,n},\ldots,\bar{T}_{k,n},T_2=\underline{T}_{q,n},\ldots,\bar{T}_{q,n}}$, and where the decreasing sequence goes from 1 to $\frac{\underline{T}_{k,n}}{\bar{T}_{k,n}}\frac{\underline{T}_{q,n}}{\bar{T}_{q,n}}$, we have

$$\begin{aligned}
&\left|\mathbb{E}\Big[\big(m_q-\mu_q\big)\big(m_k-\mu_k\big)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{\underline{T}_{q,n}}{\underline{T}_{q,n}+t'}\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\right|\\
&\leq\left|\mathbb{E}\Big[\big[\big(m_q-\mu_q\big)\big(m_k-\mu_k\big)-c\big]\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{\underline{T}_{q,n}}{\underline{T}_{q,n}+t'}\right.\\
&\left.\times\,\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\right|+r\\
&\leq\sqrt{\mathbb{E}\Big[\big(m_q-\mu_q\big)^2\mathbb{I}\left\{\xi\right\}\Big]}\sqrt{\mathbb{E}\Big[\big(m_k-\mu_k\big)^2\mathbb{I}\left\{\xi\right\}\Big]}\Big[1-\frac{\underline{T}_{k,n}}{\bar{T}_{k,n}}\frac{\underline{T}_{q,n}}{\bar{T}_{q,n}}\Big]+r\\
&\leq\frac{\sigma_k}{\sqrt{\underline{T}_{k,n}}}\frac{\sigma_q}{\sqrt{\underline{T}_{q,n}}}\Big[1-\frac{\underline{T}_{k,n}}{\bar{T}_{k,n}}\frac{\underline{T}_{q,n}}{\bar{T}_{q,n}}\Big]+r,
\end{aligned}$$

where $r$ comes from Step 2 and from the fact that $\Big|\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{\underline{T}_{q,n}}{\underline{T}_{q,n}+t'}\mathbb{I}\left\{\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big|\leq 1$.
Note that, by Step 2:

$$\begin{aligned}
&w_kw_q\Big[\frac{\sigma_k}{\sqrt{\underline{T}_{k,n}}}\frac{\sigma_q}{\sqrt{\underline{T}_{q,n}}}\Big[1-\frac{\underline{T}_{k,n}}{\bar{T}_{k,n}}\frac{\underline{T}_{q,n}}{\bar{T}_{q,n}}\Big]+r\Big]\\
&\leq\frac{\Sigma_w^2}{n^2}\frac{1}{\sqrt{\underline{T}_{k,n}}}\frac{1}{\sqrt{\underline{T}_{q,n}}}\Big[\bar{T}_{k,n}\bar{T}_{q,n}-\underline{T}_{k,n}\underline{T}_{q,n}\Big]+r\\
&=r+\frac{\Sigma_w^2}{n^2}\frac{1}{\sqrt{\underline{T}_{k,n}}}\frac{1}{\sqrt{\underline{T}_{q,n}}}\\
&\times\Big[(\bar{T}_{q,n}-\underline{T}_{q,n})\underline{T}_{k,n}+(\bar{T}_{k,n}-\underline{T}_{k,n})\underline{T}_{q,n}+(\bar{T}_{q,n}-\underline{T}_{q,n})(\bar{T}_{k,n}-\underline{T}_{k,n})\Big]+r\\
&\leq\frac{\Sigma_w^2}{n^2}\frac{1}{\lambda_{\min}n-B\sqrt{n}}\Big[3(B+D)n\sqrt{n}\Big]+r\\
&\leq\tilde{O}(n^{-3/2}),
\end{aligned}$$

because $r=O(n^{-3/2})$.

**Step 4.2: Bounding the second and third term (Equations 49 and 50).** We want to bound Equation 49

$$\Big|\mathbb{E}\Big[(m_k-\mu_k)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\Big(\frac{1}{\underline{T}_{q,n}+t'}\sum_{v=\underline{T}_{q,n}+1}^{\underline{T}_{q,n}+t'}(X_{q,v}-\mu_q)\Big)\Big]$$
$$\times\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]$$
$$=\Big|\sum_{v=\underline{T}_{q,n}+1}^{\bar{T}_{q,n}}\mathbb{E}\Big[(m_k-\mu_k)(X_{q,v}-\mu_q)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{1}{\underline{T}_{q,n}+t'}\Big]\mathbb{I}\Big\{v\leq\underline{T_{q,n}}+t'\Big\}$$
$$\times\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\Big|.$$

Let us fix $v$.

At first, note that because of Equation 46

$$\Big|\mathbb{E}\Big[(m_k-\mu_k)(X_{q,v}-\mu_q)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{1}{\underline{T}_{q,n}+t'}\Big]\mathbb{I}\Big\{v\leq\underline{T_{q,n}}+t'\Big\}$$
$$\times\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\Big|$$
$$\leq\Big|\mathbb{E}\Big[(m_k-\mu_k)(X_{q,v}-\mu_q)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{1}{\underline{T}_{q,n}+t'}\Big]\mathbb{I}\Big\{v\leq\underline{T_{q,n}}+t'\Big\}$$
$$\times\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]$$
$$+\mathbb{E}\Big[(m_k-\mu_k)(X_{q,v}-\mu_q)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{1}{\underline{T}_{q,n}+t'}\Big]\mathbb{I}\Big\{v>\underline{T_{q,n}}+t'\Big\}$$
$$\times\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\Big|+r\frac{1}{\underline{T_{q,n}}}(\bar{T}_{q,n}-\underline{T}_{q,n})$$
$$\leq\Big|\mathbb{E}\Big[(m_k-\mu_k)(X_{q,v}-\mu_q)\sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}}\sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}}\Big[\frac{\underline{T}_{k,n}}{\underline{T}_{k,n}+t}\frac{1}{\underline{T}_{q,n}+t'}\Big]$$
$$\times\mathbb{I}\left\{T_{k,n}=\underline{T}_{k,n}+t,T_{q,n}=\underline{T}_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\Big|+r\frac{1}{\underline{T_{q,n}}}(\bar{T}_{q,n}-\underline{T}_{q,n}).$$

Let us pose $c=\mathbb{E}\Big[(m_k-\mu_k)(X_{q,v}-\mu_q)\mathbb{I}\left\{\xi\right\}\Big]$.

By Lemma 6 where the partition of $\xi$ is $\Big\{\xi\cap\{T_{k,n}=T_1,T_{q,n}=T_2\}\Big\}_{T_1=\underline{T}_{k,n},\ldots,\bar{T}_{k,n},T_2=\underline{T}_{q,n},\ldots,\bar{T}_{q,n}}$, and where the decreasing sequence goes from $\frac{1}{\underline{T_{q,n}}}$ to $\frac{\underline{T}_{k,n}}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}$, we have

$$
\begin{aligned}
&\Big|\mathbb{E}\Big[\big(m_k-\mu_k\big)\big(X_{q,v}-\mu_q\big)\sum_{t=0}^{\bar T_{k,n}-\underline T_{k,n}}\sum_{t'=0}^{\bar T_{q,n}-\underline T_{q,n}}\Big[\frac{\underline T_{k,n}}{\underline T_{k,n}+t}\frac{1}{\underline T_{q,n}+t'}\Big]\\
&\times\mathbb{I}\left\{T_{k,n}=\underline T_{k,n}+t,T_{q,n}=\underline T_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\Big|\\
&\le\Big|\mathbb{E}\Big[\Big[\big(m_k-\mu_k\big)\big(X_{q,v}-\mu_q\big)-c\Big]\sum_{t=0}^{\bar T_{k,n}-\underline T_{k,n}}\sum_{t'=0}^{\bar T_{q,n}-\underline T_{q,n}}\frac{\underline T_{k,n}}{\underline T_{k,n}+t}\frac{1}{\underline T_{q,n}+t'}\\
&\times\mathbb{I}\left\{\underline T_{k,n}+t,T_{q,n}=\underline T_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\Big|+\frac{r}{\underline T_{q,n}}\\
&\le\sqrt{\mathbb{E}\Big[\big(X_{q,v}-\mu_q)^2\big)\mathbb{I}\left\{\xi\right\}\Big]}\sqrt{\mathbb{E}\Big[\big(m_k-\mu_k\big)^2\mathbb{I}\left\{\xi\right\}\Big]}\Big[\frac{1}{\underline T_{q,n}}-\frac{\underline T_{k,n}}{\bar T_{k,n}}\frac{1}{\bar T_{q,n}}\Big]+\frac{r}{\underline T_{q,n}}\\
&\le\frac{\sigma_k}{\sqrt{\underline T_{k,n}}}\sigma_q\Big[\frac{1}{\underline T_{q,n}}-\frac{\underline T_{k,n}}{\bar T_{k,n}}\frac{1}{\bar T_{q,n}}\Big]+\frac{r}{\underline T_{q,n}}.
\end{aligned}
$$

And then by summing over all $v$, we bound the second term by $(\frac{\sigma_k}{\sqrt{\underline T_{k,n}}}\sigma_q\Big[\frac{1}{\underline T_{q,n}}-\frac{\underline T_{k,n}}{\bar T_{k,n}}\frac{1}{\bar T_{q,n}}\Big]+\frac{2r}{\underline T_{q,n}})(\bar T_{q,n}-\underline T_{q,n})$.

Note that

$$
\begin{aligned}
&w_kw_q(\frac{\sigma_k}{\sqrt{\underline T_{k,n}}}\sigma_q\Big[\frac{1}{\underline T_{q,n}}-\frac{\underline T_{k,n}}{\bar T_{k,n}}\frac{1}{\bar T_{q,n}}\Big]+\frac{2r}{\underline T_{q,n}})(\bar T_{q,n}-\underline T_{q,n})\\
&\le 2r\frac{(A+C)n^{2/3}}{En^{2/3}}+\frac{\Sigma_w^2}{n^2}\frac{1}{\sqrt{\underline T_{k,n}}}\frac{1}{\underline T_{q,n}}(\bar T_{q,n}-\underline T_{q,n})\\
&\times\Big[(\bar T_{q,n}-\underline T_{q,n})\underline T_{k,n}+(\bar T_{k,n}-\underline T_{k,n})\underline T_{q,n}+(\bar T_{q,n}-\underline T_{q,n})(\bar T_{k,n}-\underline T_{k,n})\Big]\\
&\le 2r\frac{(A+C)}{E}+\frac{\Sigma_w^2}{n^2}(\frac{1}{n\lambda_{\min}-B\sqrt{n}})^{3/2}(B+D)\sqrt{n}\Big[3n(B+D)\sqrt{n}\Big]\\
&\le\tilde O(n^{-3/2}).
\end{aligned}
$$

We bound the third term (Equation 50) in the same way by $(\frac{\sigma_q}{\sqrt{\underline T_{q,n}}}\sigma_k\Big[\frac{1}{\underline T_{k,n}}-\frac{\underline T_{q,n}}{\bar T_{q,n}}\frac{1}{\bar T_{k,n}}\Big]+\frac{2r}{\underline T_{k,n}})(\bar T_{k,n}-\underline T_{k,n})$, and we can then also bound $w_kw_q$ times the third term by $\tilde O(n^{-3/2})$.

**Step 4.3: Bounding the last term (Equation 51).** We want to bound Equation 51

$$
\begin{aligned}
&\Big|\mathbb{E}\Big[\sum_{t=0}^{\bar T_{k,n}-\underline T_{k,n}}\sum_{t'=0}^{\bar T_{q,n}-\underline T_{q,n}}\Big[\big(\frac{1}{\underline T_{q,n}+t'}\sum_{v=\underline T_{q,n}+1}^{\underline T_{q,n}+t'}(X_{q,v}-\mu_q)\big)\big(\frac{1}{\underline T_{k,n}+t}\sum_{u=\underline T_{k,n}+1}^{\underline T_{k,n}+t}(X_{k,u}-\mu_k)\big)\Big]\\
&\times\mathbb{I}\left\{T_{k,n}=\underline T_{k,n}+t,T_{q,n}=\underline T_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\Big]\Big|\\
&=\Big|\sum_{v=\underline T_{q,n}+1}^{\bar T_{q,n}}\sum_{u=\underline T_{k,n}+1}^{\bar T_{k,n}}\mathbb{E}\Big[\big(X_{q,v}-\mu_q\big)\big(X_{k,u}-\mu_k\big)\sum_{t=0}^{\bar T_{k,n}-\underline T_{k,n}}\sum_{t'=0}^{\bar T_{q,n}-\underline T_{q,n}}\frac{1}{\underline T_{q,n}+t'}\frac{1}{\underline T_{k,n}+t}\\
&\times\mathbb{I}\left\{T_{k,n}=\underline T_{k,n}+t,T_{q,n}=\underline T_{q,n}+t'\right\}\mathbb{I}\left\{\xi\right\}\mathbb{I}\left\{v\le\underline T_{q,n}+t'\right\}\mathbb{I}\left\{u\le\underline T_{k,n}+t\right\}\Big]\Big|.
\end{aligned}
$$

Let us fix $v$ and $u$.

As in Step 4.2 and using 47, we obtain

$$\begin{aligned}
&\Big| \sum_{v=\underline{T}_{q,n}+1}^{\bar{T}_{q,n}} \sum_{u=\underline{T}_{k,n}+1}^{\bar{T}_{k,n}} \mathbb{E}\Big[\big(X_{q,v}-\mu_q\big)\big(X_{k,u}-\mu_k\big) \sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}} \sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}} \frac{1}{\underline{T}_{q,n}+t'} \frac{1}{\underline{T}_{k,n}+t} \\
&\times \mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t, T_{q,n}=\underline{T}_{q,n}+t'\big\}\mathbb{I}\{\xi\}\mathbb{I}\big\{v\le \underline{T}_{q,n}+t'\big\}\mathbb{I}\big\{u\le \underline{T}_{k,n}+t\big\}\Big]\Big| \\
&\le \Big| \sum_{v=\underline{T}_{q,n}+1}^{\bar{T}_{q,n}} \sum_{u=\underline{T}_{k,n}+1}^{\bar{T}_{k,n}} \mathbb{E}\Big[\big(X_{q,v}-\mu_q\big)\big(X_{k,u}-\mu_k\big) \sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}} \sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}} \frac{1}{\underline{T}_{q,n}+t'} \frac{1}{\underline{T}_{k,n}+t} \\
&\times \mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t, T_{q,n}=\underline{T}_{q,n}+t'\big\}\mathbb{I}\{\xi\}\Big]\Big| + r\frac{1}{\underline{T}_{q,n}}\frac{1}{\underline{T}_{k,n}}(\bar{T}_{q,n}-\underline{T}_{q,n})(\bar{T}_{k,n}-\underline{T}_{k,n}).
\end{aligned}$$

Let us pose $c=\mathbb{E}\Big[(X_{k,u}-\mu_k)(X_{q,v}-\mu_q)\mathbb{I}\{\xi\}\Big]$.

By Lemma 6 where the partition of $\xi$ is $\Big\{\xi\cap\{T_{k,n}=T_1,T_{q,n}=T_2\}\Big\}_{T_1=\underline{T}_{k,n},\dots,\bar{T}_{k,n},T_2=\underline{T}_{q,n},\dots,\bar{T}_{q,n}}$, and where the decreasing sequence goes from $\frac{1}{\underline{T}_{q,n}}\frac{1}{\underline{T}_{k,n}}$ to $\frac{1}{\bar{T}_{q,n}}\frac{1}{\bar{T}_{k,n}}$, we have

$$\begin{aligned}
&\Big|\mathbb{E}\Big[\big(X_{q,v}-\mu_q\big)\big(X_{k,u}-\mu_k\big) \sum_{t=0}^{\bar{T}_{k,n}-\underline{T}_{k,n}} \sum_{t'=0}^{\bar{T}_{q,n}-\underline{T}_{q,n}} \frac{1}{\underline{T}_{q,n}+t'} \frac{1}{\underline{T}_{k,n}+t} \\
&\times \mathbb{I}\big\{T_{k,n}=\underline{T}_{k,n}+t, T_{q,n}=\underline{T}_{q,n}+t'\big\}\mathbb{I}\{\xi\}\Big]\Big| \\
&\le \sqrt{\mathbb{E}\Big[\big(X_{q,v}-\mu_q\big)^2\mathbb{I}\{\xi\}\Big]}\sqrt{\mathbb{E}\Big[\big(X_{k,u}-\mu_k\big)^2\mathbb{I}\{\xi\}\Big]}\Big[\frac{1}{\underline{T}_{q,n}}\frac{1}{\underline{T}_{k,n}}-\frac{1}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}\Big]+\frac{r}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}} \\
&\le \frac{\sigma_k}{\sqrt{\underline{T}_{k,n}}}\frac{\sigma_q}{\sqrt{\underline{T}_{q,n}}}\Big[\frac{1}{\underline{T}_{q,n}}\frac{1}{\underline{T}_{k,n}}-\frac{1}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}\Big]+\frac{r}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}.
\end{aligned}$$

And then by summing over all $v$ and $u$, we bound the last term by $\Big(\frac{\sigma_k}{\sqrt{\underline{T}_{k,n}}}\frac{\sigma_q}{\sqrt{\underline{T}_{q,n}}}\Big[\frac{1}{\underline{T}_{q,n}}\frac{1}{\underline{T}_{k,n}}-\frac{1}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}\Big]+\frac{2r}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}\Big)(\bar{T}_{q,n}-\underline{T}_{q,n})(\bar{T}_{k,n}-\underline{T}_{k,n})$.

Here, also, the same way as before,

$$\begin{aligned}
&w_kw_q\Big(\frac{\sigma_k}{\sqrt{\underline{T}_{k,n}}}\frac{\sigma_q}{\sqrt{\underline{T}_{q,n}}}\Big[\frac{1}{\underline{T}_{q,n}}\frac{1}{\underline{T}_{k,n}}-\frac{1}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}\Big]+\frac{2r}{\bar{T}_{k,n}}\frac{1}{\bar{T}_{q,n}}\Big)(\bar{T}_{q,n}-\underline{T}_{q,n})(\bar{T}_{k,n}-\underline{T}_{k,n}) \\
&\le 2r(\frac{A+C}{E})^2+\frac{\Sigma_w^2}{n^2}\frac{1}{\sqrt{\underline{T}_{k,n}}}\frac{1}{\sqrt{\underline{T}_{q,n}}}\frac{1}{\underline{T}_{q,n}}\frac{1}{\underline{T}_{k,n}}(\bar{T}_{q,n}-\underline{T}_{q,n})(\bar{T}_{k,n}-\underline{T}_{k,n}) \\
&\times\Big[(\bar{T}_{q,n}-\underline{T}_{q,n})\underline{T}_{k,n}+(\bar{T}_{k,n}-\underline{T}_{k,n})\underline{T}_{q,n}+(\bar{T}_{q,n}-\underline{T}_{q,n})(\bar{T}_{k,n}-\underline{T}_{k,n})\Big] \\
&\le 2r(\frac{A+C}{E})^2+\frac{\Sigma_w^2}{n^2}(\frac{1}{n\lambda_{\min}-B\sqrt{n}})^3(B+D)n(3n(B+D)\sqrt{n}) \\
&\le \tilde{O}(n^{-3/2}).
\end{aligned}$$

**Step 5: Bounding the cross-products and finishing the proof.** By using step 4 and step 3, we get

$$w_k w_q \mathbb{E}\Big[\big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\left\{\xi\right\}\Big] \leq \tilde{O}(n^{-3/2}).$$

And

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

by using for the last inequality the result in Appendix A.

To conclude the proof, it is sufficient to use

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

□

### E.3 A distribution independent bound in the general case

**Proposition 4** *For algorithm MC-UCB when run with the parameters as in Theorem 1 and 2 , we have*

$$\mathbb{E}[(\hat{\mu}_n - \mu)^2] \leq L_n + O(n^{-7/6}).$$

*Proof:*

**Step 0: A useful Lemma.**

**Lemma 6** *Let $X$ be a random variables such that $\mathbb{E}(X) = 0$. Let $(\Omega_u)_{u=1,\ldots,p}$ be a partition of the space of random events. Let $(a_u)_{u=1,\ldots,p}$ be a positive decreasing sequence of random numbers. We have*

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

*Proof:*

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

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

This implies

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

by Cauchy-Schwartz.

By remarking that

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

we show in the same way that

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

Those two inequalities lead to the desired result.

□

**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.

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

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 $\mathcal{F}_{n,\dots,t-1,\dots,n}$[13]: 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 $\tau$ and in which one. This means that $X_{k,t}$ is independent of all events in $\tau$. Finally, we have

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

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

$$\mathbb{E}\Big[\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)\big(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\big)\Big] = 0. \tag{52}$$

Note that we have by doing a similar reasoning, that

$$\mathbb{E}\Big[\big(\sum_{t=\max(T_{k,n},\underline{T}_k)}^{\min(T_{k,n},\bar{T}_k)}(X_{k,t}-\mu_k)\big)\big(\sum_{t'=\max(T_{q,n},\underline{T}_q)}^{\min(T_{q,n},\bar{T}_q)}(X_{q,t'}-\mu_q)\big)\Big] = 0, \tag{53}$$

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

**Step 2: Definition of an event $\tau$ of high probability.** We remind that, on $\xi$ we have for all $p$,

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

and

$$T_{p,n} \leq \bar{T}_{p,n} = T^*_{p,n} + Cn^{2/3},$$

where $A$ and $C$ are as in Theorem 1 and 2 and $E$ is as in the proof of Theorem 2.

The probability of $\xi$ is more than $1-2nK\delta$.

Now let us define the event $\tau$ such that for all $p$,

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

and

$$T_{p,n} \leq \bar{T}_{p,n} = T^*_{p,n} + Cn^{2/3}.$$

Note that $\xi \subset \tau$.

We have, because of $\xi \subset \tau$,

[13]Here there are $n$ at all positions except at the $k-$ where there is a $t$.

$$
\begin{aligned}
&|\mathbb{E}[(X_{q,v}-\mu_q)(X_{k,u}-\mu_k)\mathbb{I}\{\tau^c\}]| \\
&\leq \sqrt{\mathbb{E}[(X_{q,v}-\mu_q)^2\mathbb{I}\{\tau^c\}]}\sqrt{\mathbb{E}[(X_{k,u}-\mu_k)^2\mathbb{I}\{\tau^c\}]} \\
&\leq \sqrt{\mathbb{E}[(X_{q,v}-\mu_q)^2\mathbb{I}\{\xi^c\}]}\sqrt{\mathbb{E}[(X_{k,u}-\mu_k)^2\mathbb{I}\{\xi^c\}]} \\
&\leq 2c_1nK\delta(1+\log(c_2/2nK\delta)),
\end{aligned}
$$

as in Appendix A.

Finally, because $T_{k,n}$ and $T_{q,n}$ are smaller than $n$, we have

$$
\Big|\mathbb{E}\Big[\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)\big(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\big)\mathbb{I}\{\tau^c\}\Big]\Big| \leq 2c_1n^3K\delta(1+\log(c_2/2nK\delta)),
$$

**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

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

Note now that

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

Finally, one gets

$$
\begin{aligned}
&w_k w_q \Big| \mathbb{E}\Big[ \big(\hat{\mu}_{k,n} - \mu_k\big)\big(\hat{\mu}_{q,n} - \mu_q\big)\mathbb{I}\{\tau\} \Big] \Big| \leq w_k \sigma_k \sqrt{\mathbb{E}[T_{k,n}]} w_q \sigma_q \sqrt{\mathbb{E}[T_{q,n}]} \Big( \frac{1}{\underline{T}_{k,n}} \frac{1}{\underline{T}_{q,n}} - \frac{1}{\bar{T}_{k,n}} \frac{1}{\bar{T}_{q,n}} \Big) \\
&\leq \frac{\Sigma^2}{n^2} \frac{\sqrt{\mathbb{E}[T_{k,n}]\mathbb{E}[T_{q,n}]}}{\bar{T}_{k,n}\bar{T}_{q,n}} \Big( \bar{T}_{k,n}\bar{T}_{q,n} - \underline{T}_{k,n}\underline{T}_{q,n} \Big) \\
&\leq \frac{\Sigma^2}{n^2} \frac{\sqrt{\mathbb{E}[T_{k,n}]\mathbb{E}[T_{q,n}]}}{\bar{T}_{k,n}\bar{T}_{q,n}} (A + C) \max(\bar{T}_{k,n}, \bar{T}_{q,n}) n^{2/3} \\
&\leq \frac{\Sigma^2}{n^{4/3}} (A + C) \frac{\sqrt{\mathbb{E}[T_{k,n}]\mathbb{E}[T_{q,n}]}}{\min(\bar{T}_{k,n}, \bar{T}_{q,n})} \\
&\leq \frac{\Sigma^2}{n^{7/6}} \frac{A + C}{E\sqrt{E}}.
\end{aligned}
$$

**Step 4: Finishing the proof.** Note that

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

by using for the last inequality the result in Appendix A and Step 3.

To conclude the proof, it is sufficient to use

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

□