Journal of Machine Learning Research 16 (2015) 2231-2271 Submitted 3/13; Revised 12/14; Published 8/15

Adaptive Strategy for Stratified Monte Carlo Sampling

Alexandra Carpentier A.CARPENTIER@STATSLAB.CAM.AC.UK
Statistical Laboratory
Center for Mathematical Sciences
Wilberforce Road
CB3 0WB Cambridge, United Kingdom

Remi Munos[*] MUNOS@GOOGLE.COM
Google DeepMind
London, UK

András Antos[†] ANTOS@CS.BME.HU
Budapest University of Technology and Economics
3 Műegyetem rkp.
1111 Budapest, Hungary

Editor: Nicolas Vayatis

Abstract

We consider the problem of stratified sampling for Monte Carlo integration of a random variable. We model this problem in a K-armed bandit, where the arms represent the K strata. The goal is to estimate the integral mean, that is a weighted average of the mean values of the arms. The learner is allowed to sample the variable n times, but it can decide on-line which stratum to sample next. We propose an UCB-type strategy that samples the arms according to an upper bound on their estimated standard deviations. We compare its performance to an ideal sample allocation that knows the standard deviations of the arms. For sub-Gaussian arm distributions, we provide bounds on the total regret: a distribution-dependent bound of order $\mathrm{poly}(\lambda_{\min}^{-1})\widetilde{O}(n^{-3/2})$[1] that depends on a measure of the disparity $\lambda_{\min}$ of the per stratum variances and a distribution-free bound $\mathrm{poly}(K)\widetilde{O}(n^{-7/6})$ that does not. We give similar, but somewhat sharper bounds on a proxy of the regret. The problem-independent bound for this proxy matches its recent minimax lower bound in terms of n up to a $\log n$ factor.

Keywords: adaptive sampling, bandit theory, stratified Monte Carlo, minimax strategies, active learning

1. Introduction

Estimation of mean values (or, especially, probabilities) can be considered as a special case of most problems in stochastic machine learning (e.g., regression function estimation, classification, clustering), thus understanding all of its aspects is crucial to tackle more

[*] Also affiliated to Inria Lille - Nord Europe, France

[†] During parts of this work he was with the Computer and Automation Research Institute of the Hungarian Academy of Sciences, Budapest, Hungary.

[1] The notation $a_n = \mathrm{poly}(b_n)$ means that there exist $C,\alpha > 0$ such that $a_n \le C b_n^{\alpha}$ for n large enough. Moreover, $a_n = \widetilde{O}(b_n)$ means that $a_n/b_n = \mathrm{poly}(\log n)$ for n large enough.

complex problems. 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 importance. 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 (Glasserman, 2004; Rubinstein and Kroese, 2008, Subsection 5.5). It is sometimes used in conjunction with other variance reduction techniques, such as importance sampling, antithetic sampling, or control-variables. However, in contrast with those mentioned above, stratified sampling can be used even without substantial knowledge about the function to be evaluated or the sampling distribution (though, to construct effective strata, some knowledge on the variance on different domain areas is better).

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 (Arouna, 2004; 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 paper, they propose to sample according to the empirical estimates of the standard deviations 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 research.

These works provide asymptotic convergence of the variance of the estimate to the targeted stratified variance divided by the sample size (Rubinstein and Kroese, 2008, Subsection 5.5), see also (5) in this paper. They also prove that the number of pulls within each stratum converges asymptotically to the desired number of pulls, that is, 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. For quite complete surveys on the classical bandit setting, see for example, the surveys of Cesa-Bianchi and Lugosi (2006);

Bubeck and Cesa-Bianchi (2012), and see also the seminal papers of Lai and Robbins (1985), and Auer et al. (2002). 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).

This topic has already been formalized in terms of a bandit problem in the master thesis of Grover (2009), where an algorithm named GAFS-WL (Greedy Allocation with Forced Selection - Weighted Loss) is presented. It deals with stratified sampling, that is, it targets an allocation which is proportional to the standard deviation (and not to the variance) of a stratum times its size, see the book of Rubinstein and Kroese (2008) and also as explained later on in this paper. Grover (2009) defines a proxy on the overall mean squared error (MSE, defined in Equation 1 below), the weighted sum of the per stratum MSE's (defined in Equation 3 below), that he calls *loss.* He proves that the difference between this loss of GAFS-WL and the optimal static loss is of order $\mathrm{poly}(K)\widetilde{O}(n^{-3/2})$, where the $\widetilde{O}(\cdot)$ depends of the arm distributions. Another approach for this problem, still with a bandit formalism, can be found in the paper of Carpentier and Munos (2011), where another algorithm, based on Upper-Confidence-Bounds (UCB) on the standard deviations, was proposed. This algorithm is inspired by the celebrated UCB strategy (Auer et al., 2002), that is designed for the classical bandit setting. The algorithm, called MC-UCB, samples the arms proportionally to an UCB on the standard deviation times the size of the stratum. The authors provided finite-time, problem-dependent and problem-independent bounds for the weighted MSE loss of this algorithm. The first one corresponds to the bound in the work of Grover (2009), the latter one differs from it. Finally, Carpentier and Munos (2012) developed a lower bound for this problem, stating that the pseudo-regret (defined in Section 2 below) of any algorithm for this problem cannot be significantly smaller in a problem-independent minimax sense than $\frac{K^{1/3}}{n^{4/3}}$. In addition, they prove that the problem-independent upper bound on the pseudo-regret of MC-UCB matches this bound up to some $\log n$ factor.

Note that a different, but closely analogous problem is when, instead of a weighted sum of the per arm MSE's, the maximum of these MSE's have to be minimized (e.g., because the weights are unknown). This is dealt with by Carpentier et al. (2011, 2015) for UCB-type algorithms (CH-AS, B-AS) and by Antos et al. (2010) for GAFS-type algorithm (GAFS-MAX).

Recall that in our original stratified sampling problem, however, the natural intuitive measure of performance is not the weighted MSE loss defined by Grover (2009); Carpentier and Munos (2011, 2012), but the total MSE of estimating the weighted average of the mean values of the strata. It is a very important open question to link this total MSE loss to the weighted MSE loss. Without this link, the theoretical analyses which are provided do not give bounds in terms of the natural performance measure.

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

- We provide finite-time bounds on the MSE of the estimate of the mean value. To the best of our knowledge, these are the first finite-time results for the problem of adaptive stratified Monte Carlo which target directly a usual loss measure (i.e., the total MSE). These consist of: (i) A distribution-dependent bound of order $\text{poly}(\lambda_{\min}^{-1})\widetilde{O}(n^{-3/2})$ that depends on the disparity $\lambda_{\min}$ of the strata (a measure of the problem complexity defined in Equation 6 below), and which corresponds to a stationary regime where the budget n is large compared to this complexity. (ii) A distribution-free bound of order $\text{poly}(K)\widetilde{O}(n^{-7/6})$ that does not depend on the disparity of the strata, and corresponds to a transitory regime where n is small compared to the problem complexity. (iii) The latter bound is sharpened to order $\text{poly}(K)\widetilde{O}(n^{-4/3})$ when each arm distribution is symmetric. Notably, all these bounds yield $o(1/n)$ regret rate.

- We detail the proofs of Carpentier and Munos (2011), which have not been published in full version due to space constraints. They correspond to two pseudo-regret bounds: a distribution-dependent one of order $\lambda_{\min}^{-3/2}\widetilde{O}(n^{-3/2})$ and a distribution-free one of order $K^{1/3}\widetilde{O}(n^{-4/3})$.

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 on the number of pulls, the weighted MSE loss, the total MSE loss, and the pseudo-loss under sub-Gaussian assumption on the arm distributions. We then discuss the results in Section 4. Finally, Section 5 concludes the paper and suggests future works. The appendices contain useful lemmata and the proofs.

2. Preliminaries

The allocation problem mentioned in the previous section is formalized as a K-armed bandit problem where each arm (stratum) $k = 1, \dots, K$ is characterized by a distribution ν_k with mean value μ_k and variance σ_k^2. At each round $t \geq 1$, an allocation strategy (or algorithm) $\mathcal{A}$ selects an arm k_t adaptively based on past samples, and then receives a sample drawn from ν_{k_t} that is conditionally independent of the past samples given k_t. Let $(w_k)_{k=1,\dots,K}$ denote a known set of positive weights (measure of stratum i) which sum to 1. 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 $\mathbb{I}\{E\}$ be the indicator variable of event E, that is, $\mathbb{I}\{E\} = 1$ if and only if E holds, otherwise $\mathbb{I}\{E\} = 0$. Let us write $T_{k,t} = \sum_{s=1}^t \mathbb{I}\{k_s = k\}$ for the number of times arm k has been pulled up to time t and $\hat{\mu}_{k,t} = \frac{1}{T_{k,t}} \sum_{s=1}^{T_{k,t}} X_{k,s}$ for the empirical estimate of the mean μ_k at time t, where $X_{k,s}$ denotes the sample received when pulling arm k for the s^{th} time. After n rounds, an algorithm $\mathcal{A}$ returns the empirical estimate $\hat{\mu}_{k,n}$ of μ_k for each arm and also their weighted average $\hat{\mu}_n = \sum_{k=1}^K w_k \hat{\mu}_{k,n}$ as the empirical estimate of μ.

For any algorithm $\mathcal{A}$, we use the *total mean (expected) squared error* (MSE) loss of $\hat{\mu}_n$ as performance measure in estimating μ:

$$\bar{L}_n(\mathcal{A}) = \mathbb{E}\big[(\hat{\mu}_n - \mu)^2\big] = \mathbb{E}\bigg[\Big(\sum_{k=1}^K w_k(\hat{\mu}_{k,n} - \mu_k)\Big)^2\bigg], \tag{1}$$

where $\mathbb{E}[\cdot]$ is the expectation integrated over all the samples of all arms. The goal is to define an allocation strategy that minimizes the total MSE loss defined by (1). The total

MSE loss can be decomposed as

$$\bar{L}_n(\mathcal{A}) = \underbrace{\sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big]}_{L_n(\mathcal{A})} + \sum_{k=1}^{K} \sum_{k' \neq k} w_k w_{k'} \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{k',n} - \mu_{k'})\big]. \qquad (2)$$

Here the *weighted MSE loss*

$$L_n(\mathcal{A}) = \sum_{k=1}^{K} w_k^2 \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big] \qquad (3)$$

is equal to the loss defined by Grover (2009); Carpentier and Munos (2011). Thus our analysis for stratified sampling problem implicitly covers the other problem, where instead of estimating μ, the goal is estimating all μ_k simultaneously under a weighted MSE loss $L'_n(\mathcal{A}) = \sum_{k=1}^{K} p_k(\hat{\mu}_{k,n} - \mu_k)^2$, since this loss is essentially the same as $L_n(\mathcal{A})$. Such a setting is referred to sometimes as an *active learning* (or active regression estimation) problem in the literature (e.g., Grover, 2009). This case is even simpler in the sense that we do not have to bother with the cross product-terms in (2).

Note that if all the $T_{k,n}$ are *deterministic*, then in the cross product-terms

$$\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{k',n} - \mu_{k'})\big] = \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)\big]\mathbb{E}\big[(\hat{\mu}_{k',n} - \mu_{k'})\big] = 0 \cdot 0 = 0,$$

and also $\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big] = \sigma_k^2 / T_{k,n}$. This implies that in this case

$$\bar{L}_n(\mathcal{A}) = L_n(\mathcal{A}) = \sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2}{T_{k,n}}. \qquad (4)$$

This gives rise to the definition of

$$\widetilde{L}_n(\mathcal{A}) = \sum_{k=1}^{K} w_k^2 \mathbb{E}\left[\frac{\sigma_k^2}{T_{k,n}}\right]$$

for any algorithm $\mathcal{A}$ (with sample dependent $T_{k,n}$'s) as an alternative performance measure. We call $\widetilde{L}_n(\mathcal{A})$ *pseudo-loss*, as it is a proxy of $\bar{L}_n(\mathcal{A})$ and $L_n(\mathcal{A})$. It is obviously equal to them for deterministic $T_{k,n}$'s.

2.1 Optimal Allocation

Although (4) does not hold when the numbers of pulls of an adaptive algorithm depend on the observed samples and thus are random, it holds when each arm is pulled a deterministic number of times. Thus if the variances of the arms were known in advance, one could design an optimal deterministic (i.e., static, non-adaptive) allocation strategy $\mathcal{A}^*$ by choosing $T_{k,n} = T^*_{k,n}$ such that they minimize $\bar{L}_n$ under the constraint $\sum_{k=1}^{K} T^*_{k,n} = n$. This optimal deterministic allocation of $\mathcal{A}^*$ is to pull each arm k proportionally to $w_k\sigma_k$ (up to rounding effects), that is, given by

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

This achieves the loss

$$\bar{L}_n(\mathcal{A}^*) = L_n(\mathcal{A}^*) = \widetilde{L}_n(\mathcal{A}^*) = \frac{\Sigma_w^2}{n}, \tag{5}$$

where $\Sigma_w \overset{\text{def}}{=} \sum_{k=1}^K w_k\sigma_k$. We assume in the sequel that $\Sigma_w > 0$, that is, $\exists k$ that $\sigma_k > 0$. We define also $\bar{\Sigma} \overset{\text{def}}{=} \max_k \sigma_k$. In the following, we write

$$\lambda_k \overset{\text{def}}{=} \frac{T^*_{k,n}}{n} = \frac{w_k\sigma_k}{\Sigma_w}$$

for the optimal allocation proportion for arm k and

$$\lambda_{\min} \overset{\text{def}}{=} \min_{1\le k\le K} \lambda_k \qquad , \qquad \underline{w} \overset{\text{def}}{=} \min_{1\le k\le K} w_k. \tag{6}$$

Note that a small $\lambda_{\min}$ means a large disparity of the quantities $\{w_k\sigma_k\}_{k\le K}$. It will turn out that $\lambda_{\min}$ seems to characterize the hardness of a problem.

2.2 Uniform Allocation

Another possible deterministic allocation is the *proportional* or *uniform strategy* $\mathcal{A}^u$ which assumes uniform standard deviations (e.g., since the σ_k's are unknown and thus the optimal allocation is out of reach), that is, allocates such that $T_k^u = \frac{w_k}{\sum_{i=1}^K w_i} n = w_k n$. Its loss is

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

where $\Sigma_{w,2} = \sum_{k=1}^K w_k\sigma_k^2$. Note that using either Jensen's or Cauchy-Schwarz's inequality, we can see that $\Sigma_w^2 \le \Sigma_{w,2}$ with equality if and only if all the σ_k's are equal. Thus $\mathcal{A}^*$ is always at least as good as $\mathcal{A}^u$. In addition, since $\sum_k w_k = 1$, we have

$$\Sigma_{w,2} - \Sigma_w^2 = \sum_k w_k(\sigma_k - \Sigma_w)^2.$$

The difference between those two quantities is the weighted quadratic variation of the σ_k's ($1 \le k \le K$) around their weighted mean Σ_w. As a result the gain of $\mathcal{A}^*$ compared to $\mathcal{A}^u$ grows with the disparity of the σ_k's.

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

2.3 Definition of Regret

In order to assess how well $\mathcal{A}$ solves the *exploration-exploitation trade-off* above 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 *total/weighted MSE regret* of an adaptive algorithm $\mathcal{A}$ as the difference between the total/weighted MSE loss incurred by $\mathcal{A}$ and the optimal loss, respectively:

$$\bar{R}_n(\mathcal{A}) = \bar{L}_n(\mathcal{A}) - \frac{\Sigma_w^2}{n} \qquad , \qquad R_n(\mathcal{A}) = L_n(\mathcal{A}) - \frac{\Sigma_w^2}{n}.$$

The total MSE regret indicates how much we loose in terms of MSE by not knowing in advance the standard deviations σ_k. Note that since $\bar{L}_n(\mathcal{A}^*) \propto 1/n$ by (5), a consistent strategy, that is, one which is asymptotically equivalent to the optimal strategy, is obtained whenever its regret is negligible compared to $1/n$.

We also define the *pseudo-regret*, a proxy for the MSE regret, as the difference between the pseudo-loss incurred by the algorithm and the optimal loss:

$$\widetilde{R}_n(\mathcal{A}) = \widetilde{L}_n(\mathcal{A}) - \frac{\Sigma_w^2}{n}.$$

It is important to derive bounds for $\bar{R}_n(\mathcal{A})$ when $T_{k,n}$'s are random. Taking the decomposition (2), a natural way to proceed is to prove that both
(i) $R_n(\mathcal{A})$ is small and
(ii) the cross product-terms $\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{k',n} - \mu_{k'})\big]$ are small.

Note that for $K = 1$, for any $\mathcal{A}$, $T_{1,n} = T^*_{1,n} = n$ and $\bar{R}_n(\mathcal{A}) = R_n(\mathcal{A}) = \widetilde{R}_n(\mathcal{A}) = 0$, thus we assume $K \geq 2$ from now on.

3. Allocation Based on Monte Carlo Upper Confidence Bound

We now describe the main algorithm and the associated bounds.

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, $\mathcal{A}_{\text{MC-UCB}}$, is described in Figure 1. It requires a parameter β as input, which should be chosen as explained below after Assumption 1.

```
Input: $\beta$
Initialize: Pull each arm twice.
for $t = 2K+1, \ldots, n$ do
  Compute $B_{k,t}$ using (7) for each arm $1 \leq k \leq K$
  Pull an arm $k_t \in \arg\max_{1 \leq k \leq K} B_{k,t}$
end for
Output: $\hat{\mu}_{k,n}$ for each arm $1 \leq k \leq K$ and $\hat{\mu}_n$
```

Figure 1: The pseudo-code of the MC-UCB algorithm.

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

$$B_{k,t} = \frac{w_k}{T_{k,t-1}}\Big(\hat{\sigma}_{k,t-1} + \frac{2\beta}{\sqrt{T_{k,t-1}}}\Big) \tag{7}$$

on the standard deviation σ_k for each arm k, and then pulls the one with largest $B_{k,t}$. The bounds $B_{k,t}$ are built by using Lemma 10 (and Corollary 16) and based on the empirical standard deviation $\hat{\sigma}_{k,t-1}$:

$$\hat{\sigma}^2_{k,t-1} = \frac{1}{T_{k,t-1} - 1}\sum_{i=1}^{T_{k,t-1}} (X_{k,i} - \hat{\mu}_{k,t-1})^2, \tag{8}$$

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, $\mathcal{A}_{\text{MC-UCB}}$ returns the empirical mean $\hat{\mu}_{k,n}$ for each arm $1 \leq k \leq K$ and also their weighted average $\hat{\mu}_n$.

The motivation to use such an adaptive algorithm instead of classical strategies using, for example, a limited pre-run to get preliminary estimates of the variances is that the latter needs to know the sample size in advance, and will not be able to adapt the length of the exploration phase to the difficulty of the problem. For instance, a strategy that uses e.g., $\approx n^{2/3}$ samples for variance estimation will have minimax-optimal problem-independent rate (up to a log factor) but will display a suboptimal problem-dependent regret rate, i.e., $n^{-4/3}$. On the other hand, a strategy that uses e.g., $\approx n/\log n$ samples for variance estimation will have an optimal problem-dependent regret (of order $n^{-3/2}$ up to a log factor). The main advantage of adaptive strategies such as the one we provide is that it adapts the length of exploration phase to the difficulty of the problem.

We are giving two analyses of $\mathcal{A}_{\text{MC-UCB}}$, a *problem-dependent* and a *problem-independent* one, which are interesting in the stationary and the transitory regimes of the run time of the algorithm, respectively. We will comment on this later in Section 4.

3.2 Assumption on the Arm Distributions and Setting β

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

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

$$\mathbb{P}_{X\sim\nu_k}(|X - \mu_k| \geq \epsilon) \leq c_2 \exp(-\epsilon^2/c_1). \tag{9}$$

The parameters c_1 and c_2 characterize the maximal heaviness of the tails of the arm distributions. Since (9) is equivalent to

$$\mathbb{P}\left(|X_{k,t} - \mu_k| \geq \sqrt{c_1 \log(c_2/\delta)}\right) \leq \delta \qquad \text{for any } 0 < \delta < c_2,$$

$\sqrt{c_1 \log(c_2/\delta)}$ can be seen as a high probability bound on the centered samples.

For bounded arm distributions, parameter β of $\mathcal{A}_{\text{MC-UCB}}$ should be generally set as $c\sqrt{\log(2/\delta)}$, where c is the maximum range of the distributions and δ is a chosen significance level corresponding to the estimation of the standard deviations (see Theorem 12). In particular, δ will be chosen as an appropriate decreasing function of n (here $n^{-9/2}$) giving $\beta = \beta_n \propto c\sqrt{\log n}$.

For unbounded distributions satisfying Assumption 1, the role of c is taken by $\propto \sqrt{c_1 \log(c_2/\delta)}$, and the expressions become more involved. Then β will be set as the following function of c_1, c_2, δ, and the total sample size n

$$\beta = \beta_n(\delta) \overset{\text{def}}{=} 2\sqrt{c_1 \log(c_2/\delta)\log(2/\delta)} + \frac{\sqrt{c_1 n\delta \log(ec_2/\delta)}}{2(1-\delta)}. \tag{10}$$

This particular form comes from the way we extend a tail inequality for sub-Gaussian random variables in Proposition 14 of Appendix B. In particular, substituting $\delta = n^{-9/2}$ into (10) $\beta = \beta_n$ will be set as the following function of n, c_1, and c_2

$$\beta_n \overset{\text{def}}{=} \sqrt{c_1 \log(c_2^2 n^9)\log(4n^9)} + \frac{\sqrt{c_1 \log(ec_2 n^{4.5})}}{2(1-n^{-4.5})n^{7/4}}. \tag{11}$$

To help the reader, subscript n will be used after this substitution. Moreover, note that $B_{k,t}$, k_t, $T_{k,t}$, $\hat{\mu}_{k,t}$, and $\hat{\sigma}_{k,t}$, beside depending on the time step $t \le n$, depend, possibly in an indirect way, also on β, and so on δ, the budget n, c_1, and c_2. An accurate notation would denote also these in some indices to avoid confusion. However, since we consider mostly fixed n, δ, c_1, and c_2, we keep the lighter notations above for the sake of concision.

3.3 High-Probability Bounds on the Number of Pulls

For $2 \le t \le n$, $1 \le k \le K$, write

$$\hat{s}_{k,t}^2 \overset{\text{def}}{=} \frac{1}{t-1}\sum_{i=1}^{t}\Big(X_{k,i} - \frac{1}{t}\sum_{t'=1}^{t} X_{k,t'}\Big)^2 \tag{12}$$

for the unbiased empirical variances corresponding to the first t samples from arm k and also $\hat{s}_{k,t} \overset{\text{def}}{=} \sqrt{\hat{s}_{k,t}^2}$. Then we have $\hat{\sigma}_{k,t} = \hat{s}_{k,T_{k,t}}$ as computed in (8).

To conduct our analysis, first we state upper and lower bounds on the difference between the allocation $T_{k,n}$ implemented by the MC-UCB algorithm run by parameter β and the optimal allocation $T_{k,n}^*$ for each arm which hold on the event that all standard deviation estimations $\hat{s}_{k,t}$ are quite accurate, namely on

$$\xi = \xi_{K,n}(\delta) \overset{\text{def}}{=} \bigcap_{1\le k\le K,\ 2\le t\le n} \left\{ |\hat{s}_{k,t} - \sigma_k| \le \frac{2\beta}{\sqrt{t}} \right\}, \tag{13}$$

where β is given by (10). Later Corollary 16 will show that a small δ implies a high probability $\mathbb{P}(\xi)$ under Assumption 1, thus we can use these results to derive the various regret bounds in Subsections 3.4–3.7 for the algorithm. The proofs of Lemma 1 and 2 are in Appendix A.

Problem-dependent bound. All of our problem-dependent bounds (Lemma 1, Propositions 3, 8, partially Proposition 6 and Theorem 7) contain $1/\lambda_{\min}$ and so become void (actually trivial) if $\lambda_{\min} = 0$.[2] Thus we assume $\lambda_{\min} > 0$ in their proofs.

Lemma 1 *Let Assumption 1 hold. For any* $0 < \delta \leq 1$*,* $n \geq 4K$*, and any arm* $1 \leq p \leq K$*, on* ξ*, the allocation* $T_{p,n}$ *implemented by* $\mathcal{A}_{\textit{MC-UCB}}$ *satisfies*

$$\frac{w_p \sigma_p}{T_{p,n}} \leq \frac{\Sigma_w}{n} + \frac{12\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}, \tag{14}$$

and consequently $T_{p,n} - T^*_{p,n}$ *satisfies*

$$-4\lambda_p \left(\frac{3\beta}{\Sigma_w \lambda_{\min}^{3/2}} \sqrt{n} + K \right) \leq T_{p,n} - T^*_{p,n} \leq 4 \left(\frac{3\beta}{\Sigma_w \lambda_{\min}^{3/2}} \sqrt{n} + K \right), \tag{15}$$

where β *is given by* (10)*.*

In (15), $|T_{p,n} - T^*_{p,n}|$ is bounded by a quantity of order $\sqrt{n}$. This is directly linked to the parametric rate of convergence of the estimation of σ_k, which is of order $1/\sqrt{n}$. Note that (15) also shows the inverse dependency on the smallest optimal allocation proportion $\lambda_{\min}$.

Problem-independent bound.

Lemma 2 *Let Assumption 1 hold. For any* $0 < \delta \leq 1$*,* $n \geq 4K$*, and any arm* $1 \leq p \leq K$*, on* ξ*, the allocation* $T_{p,n}$ *implemented by* $\mathcal{A}_{\textit{MC-UCB}}$ *satisfies*

$$T_{p,n} \geq \frac{(w_p n)^{2/3}}{\gamma^2} \qquad \textit{and} \tag{16}$$

$$\frac{w_p \sigma_p}{T_{p,n}} \leq \frac{\Sigma_w}{n} + \frac{12K^{1/3}\beta\gamma}{n^{4/3}} + \frac{4K\Sigma_w}{n^2}, \tag{17}$$

and consequently $T_{p,n} - T^*_{p,n}$ *satisfies*

$$-4\lambda_p \left(\frac{3K^{1/3}\beta\gamma}{\Sigma_w} n^{2/3} + K \right) \leq T_{p,n} - T^*_{p,n} \leq 4 \left(\frac{3K^{1/3}\beta\gamma}{\Sigma_w} n^{2/3} + K \right),$$

where β *is given by* (10) *and* $\gamma = \gamma_n(\delta) \overset{\text{def}}{=} (\bar{\Sigma}/\beta + \sqrt{8})^{1/3}$*.*

Unlike in the bounds proved in Lemma 1, here $|T_{p,n} - T^*_{p,n}|$ is bounded by a quantity of order $n^{2/3}$ without any inverse dependency on $\lambda_{\min}$.

[2] There are good chances in this case that by refined analyses and setting $\lambda_{\min} = \min_{1 \leq k \leq K : \lambda_k > 0} \lambda_k$ (that is > 0), the same formulae can be proven giving finite bounds.

3.4 Bounds on the Weighted MSE Regret of $\mathcal{A}_{\textbf{MC-UCB}}$

To simplify our bounds, we introduce

$$C_\beta = C_{\beta,n} \overset{\text{def}}{=} \sqrt{c_1}(9\log n + 1.6\log(c_2+1)) \qquad \text{and} \tag{18}$$

$$C_\xi = C_{\xi,n} \overset{\text{def}}{=} c_1 \log(ec_2 n^{7/2}/2K) \tag{19}$$
$$\left(\ < c_1(7\log n/2 + \log c_2) \qquad \text{for } K \geq 2\ \right),$$

which depend only polynomially on $\log n$, $\sqrt{c_1}$, and $\log c_2$. We now report the bounds on $R_n(\mathcal{A}_{\text{MC-UCB}})$. The proofs are given in Appendix D.

Problem-dependent bound. This result depends crucially on $\lambda_{\min}^{-1}$ which is a measure of the disparity of the products of the standard deviations and the weights. For this reason we refer to it as "distribution-dependent" result. Its proof relies on the upper- and lower bounds on $T_{k,t} - T^*_{k,t}$ in Lemma 1.

Proposition 3 *Let Assumption 1 be verified for two parameters c_1, $c_2 \geq 1$. If β_n is given by* (11)*, then for $n \geq 4K$ it holds for $\mathcal{A}_{\textit{MC-UCB}}$ that*

$$R_n(\mathcal{A}_{\textit{MC-UCB}}) \leq \frac{24\Sigma_w C_\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{288 C_\beta^2}{n^2\lambda_{\min}^3} + \frac{\sqrt{K}C_\xi + 32K\Sigma_w^2}{2n^2},$$

where C_β and C_ξ are given by (18) *and* (19).

Problem-independent bound. Now we report our second bound on $R_n(\mathcal{A}_{\text{MC-UCB}})$ that does not depend on $\lambda_{\min}^{-1}$ at all. This is obtained at the price of the worse rate $K^{1/3}\widetilde{O}(n^{-4/3})$. Its proof relies on the upper- and lower bounds on $T_{k,t} - T^*_{k,t}$ in Lemma 2.

Proposition 4 *Let Assumption 1 be verified for two parameters c_1, $c_2 \geq 1$. If β_n is given by* (11)*, then for $n \geq 4K$ it holds for $\mathcal{A}_{\textit{MC-UCB}}$ that*

$$R_n(\mathcal{A}_{\textit{MC-UCB}}) \leq \frac{36K^{1/3}\Sigma_w C_\beta}{n^{4/3}} + \frac{K^{2/3}(2058C_\beta^2 + 32\Sigma_w^2) + K^{1/6}C_\xi}{(2n)^{5/3}},$$

where C_β and C_ξ are given by (18) *and* (19).

Note that this bound is not entirely distribution free, since Σ_w appears. But, as proven in Appendix B.3 using Assumption 1, $\Sigma_w^2 \leq c_1\log(ec_2)$.

For Gaussian distributions with variance 1, we can take $c_1 = c_2 = 1$, and the main coefficient of $\log n/(n\lambda_{\min})^{3/2}$ in Proposition 3 and of $K^{1/3}\log n/n^{4/3}$ in Proposition 4 are upper bounded by 216 and 324, respectively.

3.5 Bounds on the Cross Product-Terms

The difficulty in bounding the cross product-terms, that is, the second term in the right-hand side of (2), comes from the fact that the $(T_{k,n})_{k\leq K}$ depend on the samples (in particular, for $\mathcal{A}_{\text{MC-UCB}}$, on the empirical standard deviations $(\hat{\sigma}_{k,t})_{k\leq K, t\leq n}$). This dependence can make correlation between $\hat{\mu}_{k,n}$ and $\hat{\mu}_{k',n}$. Thus, for general distributions, we cannot see obvious, direct reason why a cross product-term should be equal to the product of the corresponding biases, and so be close to 0. We give three results for these cross product-terms. The first one corresponds to the specific case where the distributions of the arms are symmetric. The next two provide a problem-dependent and a problem-independent bound in the general case. All these are partial results for proving bounds on $\bar{R}_n(\mathcal{A}_{\text{MC-UCB}})$ and proven in Appendix E.

Arms with symmetric distributions. The first result holds in the specific case of symmetric distributions. Intuitively speaking, in this setting, conditioning on the empirical standard deviations does not change the mean of the samples (and sample averages). This implies that for $k \neq k'$, $\hat{\mu}_{k,n} - \mu_k$ and $\hat{\mu}_{k',n} - \mu_{k'}$ are conditionally uncorrelated. From that we deduce the following result.

Proposition 5 *Assume that each distribution ν_k is symmetric around μ_k, respectively. For $\mathcal{A}_{\text{MC-UCB}}$ launched with any parameter β_n, we have that*

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

Though mostly of theoretical interest, the significance of this result is its indication that the rate might be improvable for other distributions, as well.

Problem-dependent and problem-independent bound in general. The following proposition gives bounds on the cross product-terms. This can be seen as an intermediary step in linking the weighted MSE regret and the true regret. Its proof relies on the specific structure of $\mathcal{A}_{\text{MC-UCB}}$ through the use of Lemma 1 and 2.

Proposition 6 *Let Assumption 1 be verified for two parameters c_1, $c_2 \geq 1$. If β_n is given by* (11), *then (for n large enough compared to K, c_1, $\log c_2$, and $1/\Sigma_w$) the cross product-terms for $\mathcal{A}_{\text{MC-UCB}}$ are bounded as*

$$\sum_{k=1}^{K} \sum_{q \neq k} w_k w_q \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\big] \leq \text{poly}(\Sigma_w c_1 \log c_2/\lambda_{\min}) \widetilde{O}(n^{-3/2}),$$

and

$$\sum_{k=1}^{K} \sum_{q \neq k} w_k w_q \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\big] \leq \text{poly}(K \Sigma_w c_1 \log c_2/\underline{w}) \widetilde{O}(n^{-7/6}),$$

where $\underline{w}$ is given by (6) *(and $\widetilde{O}(\cdot)$* does not *depend on $\lambda_{\min}$).*

Note that the latter bound, depending on $\underline{w}$, is not really problem-independent (considering w_k's to be part of the problem), but it is independent of the arm distributions, particularly of $\lambda_{\min}$.

3.6 Bounds on the Total-Regret

From the decomposition (2) for $\mathcal{A}_{\text{MC-UCB}}$ and Propositions 3, 4, 6, and 5, we can deduce our main result, a bound on the true regret $\bar{R}_n(\mathcal{A}_{\text{MC-UCB}})$:

Theorem 7 *Let Assumption 1 be verified for two parameters $c_1 > 0$, $c_2 \geq 1$. If β_n is given by* (11)*, then (for n large enough compared to K, c_1, $\log c_2$, and $1/\Sigma_w$) the true regret of $\mathcal{A}_{\textit{MC-UCB}}$ is bounded as*

$$\bar{R}_n(\mathcal{A}_{\textit{MC-UCB}}) = \text{poly}(\Sigma_w c_1 \log c_2/\lambda_{\min})\widetilde{O}(n^{-3/2}),$$

and

$$\bar{R}_n(\mathcal{A}_{\textit{MC-UCB}}) = \text{poly}(K\Sigma_w c_1 \log c_2/\underline{w})\widetilde{O}(n^{-7/6})$$

(thus, in particular, $\bar{R}_n = o(1/n)$). If each distribution ν_k is symmetric around μ_k, then the cross product-terms are 0*, and the following tighter problem-independent bound holds*

$$\bar{R}_n(\mathcal{A}_{\textit{MC-UCB}}) = R_n(\mathcal{A}_{\textit{MC-UCB}}) = \text{poly}(K\Sigma_w c_1 \log c_2)\widetilde{O}(n^{-4/3}).$$

3.7 Bounds on the Pseudo-Regret

We bound $\widetilde{R}_n(\mathcal{A}_{\text{MC-UCB}})$ by a problem-dependent and a problem-independent upper bound that are of the same order in n as the bounds in Propositions 3 and 4, respectively. The proofs are given in Appendix C.

Problem-dependent bound.

Proposition 8 *Let Assumption 1 be verified for two parameters $c_1 > 0$, $c_2 \geq 1$. If β_n is given by* (11)*, then the pseudo-regret of $\mathcal{A}_{\textit{MC-UCB}}$ launched with $n \geq 4K$ is bounded as*

$$\widetilde{R}_n(\mathcal{A}_{\textit{MC-UCB}}) \leq \frac{12\Sigma_w C_\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{(4K + \sqrt{2}/16)\Sigma_w^2}{n^2},$$

where C_β is given by (18)*.*

Problem-independent bound.

Proposition 9 *Let Assumption 1 be verified for two parameters $c_1 > 0$, $c_2 \geq 1$. If β_n is given by* (11)*, then the pseudo-regret of $\mathcal{A}_{\textit{MC-UCB}}$ launched with $n \geq 4K$ is bounded as*

$$\widetilde{R}_n(\mathcal{A}_{\textit{MC-UCB}}) \leq \frac{18K^{1/3}\Sigma_w C_\beta}{n^{4/3}} + \frac{(4K + \sqrt{2}/16)\Sigma_w^2}{n^2},$$

where C_β is given by (18)*.*

For Gaussian distributions with variance 1, we can consider $c_1 = c_2 = 1$, and the main coefficient of $\log n/(n\lambda_{\min})^{3/2}$ in Proposition 8 and of $K^{1/3}\log n/n^{4/3}$ in Proposition 9 are upper bounded by 108 and 162, respectively.

4. Discussion on the Results

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

4.1 Problem-Dependent and -Independent Bounds on $R_n(\mathcal{A})$ and $\widetilde{R}_n(\mathcal{A})$

Our problem-dependent $\lambda_{\min}^{-3}\widetilde{O}(n^{-3/2})$ upper bound on $R_n(\mathcal{A}_{\text{MC-UCB}})$ in Proposition 3 is similar and comparable to the one provided for GAFS-WL by Grover (2009), where the loss measure is $L_n(\mathcal{A}_{\text{GAFS-WL}})$. Beside this $\lambda_{\min}$-dependent bound for $\mathcal{A}_{\text{MC-UCB}}$, Propositions 4 gives a $\lambda_{\min}$-independent bound of order $K^{1/3}\widetilde{O}(n^{-4/3})$. (Note however, that when there is an arm with 0 variance, GAFS-WL is likely to perform better than MC-UCB, as it will only sample this arm $O(\sqrt{n})$ times, while MC-UCB usually samples it $\Omega(n^{2/3})$ times.) Similarly, Proposition 8 provides a pseudo-regret bound of order $\lambda_{\min}^{-3/2}\widetilde{O}(n^{-3/2})$, whereas Proposition 9 gives a $\lambda_{\min}$-independently bound of order $K^{1/3}\widetilde{O}(n^{-4/3})$.

Hence, for a given problem, that is, a given $\lambda_{\min}$, the distribution-free results of Proposition 4 and 9 are more informative than the distribution-dependent results of Proposition 3 and 8, respectively, in the *transitory regime*, that is, when n is small compared to $\lambda_{\min}^{-1}$. Proposition 3 and 8 is better in the *stationary regime*, that is, for n large enough. This distinction reminds us of the difference between distribution-dependent and distribution-free bounds for the UCB algorithm in usual multi-armed bandits. In that setting, the distribution dependent bound is in $O(K \log n/\Delta)$, where Δ is the difference between the mean value of the two best arms, and the distribution-free bound is in $O(\sqrt{Kn})$ as explained by Auer et al. (2002); Audibert and Bubeck (2009). In many works, these two types of results are called *individual* and *uniform* bounds. For several models, the two bounds correspond with each other, at least in their convergence rates in the sample size for the best possible algorithms (i.e., in some minimax sense). See the thesis of Antos (1999) for a discussion. Our results and proofs suggest that our stratified sampling model is another interesting exception, where these two types of rates must be different.

At first sight, the problem of Monte Carlo integration seems to be more related to the problem of *pure exploration* (Bubeck et al., 2011; Audibert et al., 2010) than to the usual bandit setting: indeed, similarly to the setting of pure exploration, an intermediate objective (linked to the overall objective) is to allocate the number of pulls of the arms proportionally to some unknown problem-dependent quantities. However, we believe that our problem is actually more related to the standard bandit problem, since it gives rise to an exploration-exploitation trade-off.

4.2 The Parameter β of the Algorithm

We saw in (11) that the parameter β_n of $\mathcal{A}_{\text{MC-UCB}}$ should depend on n, c_1, c_2. It is actually such that $\beta_n \approx c' \log n$, where c' can be interpreted as a high probability bound on the range of the samples. We thus simply require a rough bound on the magnitude of the samples. As we saw, in the case of bounded distributions, β_n can be chosen as $\beta_n = c\sqrt{5 \log n}$, where c is a true bound on the range of the variables. This is easy to deduce by comparing Corollary 13 and Proposition 14 in Appendix B. The interpretation of this parameter β is quite similar to the interpretation of the parameter in the UCB algorithm of Auer et al. (2002), and its order of magnitude is roughly the same. (In that paper, it is assumed that the distributions of the arms are bounded.) On the other hand, the interpretation of this quantity is quite different from the interpretation of the parameter a of the algorithm UCB-E of Audibert et al. (2010), which characterizes here the complexity of the problem. This is yet another illustration from the fact that this problem is somehow more related to the standard bandit problem than to the problem of pure exploration.

4.3 Finite-Time Bounds for $\bar{R}_n(\mathcal{A}_{\textbf{MC-UCB}})$ and Asymptotic Optimality

The first result in Theorem 7 states that $\bar{R}_n(\mathcal{A}_{\text{MC-UCB}})$ is of order $\text{poly}(\lambda_{\min}^{-1})\widetilde{O}(n^{-3/2})$. This corresponds to the $\lambda_{\min}$-dependent bound on $R_n(\mathcal{A}_{\text{MC-UCB}})$. Theorem 7 also states that an upper bound on $\bar{R}_n(\mathcal{A}_{\text{MC-UCB}})$ is of order $\text{poly}(K)\widetilde{O}(n^{-7/6})$. This corresponds to the $\lambda_{\min}$-independent bound on $R_n(\mathcal{A}_{\text{MC-UCB}})$. Unfortunately, in this case, we do not obtain the same order for $\bar{R}_n(\mathcal{A}_{\text{MC-UCB}})$ as for $R_n(\mathcal{A}_{\text{MC-UCB}})$, that is, $\text{poly}(K)\widetilde{O}(n^{-4/3})$. This comes from the fact that the bound on the cross product-terms in Proposition 6 is of order $\text{poly}(K/\underline{w})\widetilde{O}(n^{-7/6})$. Whether this bound is tight or not is an open problem.

As we bound $\bar{R}_n(\mathcal{A}_{\text{MC-UCB}})$ as $o(1/n)$, $\bar{L}_n(\mathcal{A}_{\text{MC-UCB}})$ is asymptotically not more than $\bar{L}_n(\mathcal{A}^*) = \Sigma_w^2/n$ for *any* problem satisfying Assumption 1. This can be said as $\mathcal{A}_{\text{MC-UCB}}$ is (weakly) consistent; just like the algorithms of Kawai (2010); Etoré and Jourdain (2010).

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

4.4 Pseudo-Regret of $\mathcal{A}_{\textbf{MC-UCB}}$ and the Lower Bound

Carpentier and Munos (2012) provided a $\lambda_{\min}$-independent minimax lower bound for $\widetilde{R}_n(\mathcal{A})$ that is of order $K^{1/3}\Omega(n^{-4/3})$. An important achievement is that the $\lambda_{\min}$-independent upper bound on $\widetilde{R}_n(\mathcal{A}_{\text{MC-UCB}})$ in Proposition 9 is of the same order up to a logarithmic factor. Thus, regarding $\widetilde{R}_n(\mathcal{A})$, it is impossible to improve this strategy uniformly for every sub-Gaussian problem more than by a log factor.

Although we do not have a $\lambda_{\min}$-dependent lower bound on $\widetilde{R}_n(\mathcal{A})$ yet, we believe that the $\widetilde{O}(n^{-3/2})$ rate of Proposition 8 cannot be improved in n for general distributions. As it seems from the proofs in Appendix A and C, this rate is a direct consequence of the high probability bounds on the estimates of the standard deviations of the arms which are in $O(1/\sqrt{n})$, and *those bounds are tight.* Because of the minimax lower bound that is of order $K^{1/3}\Omega(n^{-4/3})$, it is however clear that there exists no algorithm with a regret of order $n^{-3/2}$ without any dependence on $\lambda_{\min}^{-1}$ (or another related problem-dependent quantity).

4.5 Making $\mathcal{A}_{\textbf{MC-UCB}}$ Anytime

An interesting question is whether and how it is possible to make $\mathcal{A}_{\text{MC-UCB}}$ anytime, that is, not requiring the knowledge of the sample horizon n in advance. Although we will not provide formal proofs of this result in this paper, we believe that setting a δ that evolves with the current time as $\delta_t = t^{-9/2}$, is sufficient to make all the regret bounds of this paper hold with slightly modified constants. Some ideas on how to prove these results can be found in the literature (Grover, 2009; Antos et al., 2010; Auer et al., 2002).

4.6 Domains of Application

Monte Carlo integration has many relevant applications in machine learning. Being able to compute precisely an integral is a prerequisite in many methods or algorithms in this field.

Some examples of possible application of the stratified Monte Carlo technique are listed below.

- There are more and more applications in machine learning that are targeting the allocation and placement of various kinds of sensors (as e.g., pollution sensors, temperature sensors, cameras of various kinds, network sensors, etc.). It is a challenge to find a way to place them efficiently, or choose at which frequency to observe their output. The placement of these sensors should depend of the objective that they have to fulfill. In some cases, one wants to use these sensors to compute an integral (for instance, the average pollution level or temperature in a region, the average amount of traffic at a certain time, the average number of customers in a given place in a supermarket, or the average amount of exchange in a network, etc.). The approach of this paper can be used to decide adaptively how to place these sensors, how frequently to inspect them, or how many of them to put depending on the area. In some other cases, the objective is that the sensors provide a good estimate of what they measure in each zone (e.g., local water pressure on a dyke). As mentioned earlier, our algorithm minimizes, with respect to the sample allocation, a weighted (over the strata) mean squared error of estimations. Therefore, our approach also provides good results in such a setting where the objective is to estimate the mean value in each zone, rather than an overall integral.

- A huge domain that is commonly handled in the machine learning community, and in which the aim is often to compute precisely integrals is Bayesian methodology. Indeed, expectations under the posterior distribution are often good estimators for some relevant parameters of the model. Being able to compute these expectations (which are well defined integrals) fast and precisely is both desirable and challenging, and our method provides an alternative for MCMC methods in the computation of such integrals.

- There are many applications in mathematical finance, for example, in the domain of pricing (which essentially sums up to the computation of a complex stochastic integral).

As mentioned below (3), omitting the cross product-terms and focusing on the weighted MSE loss our setting can be interpreted as an active learning framework. This can be a suitable model also in production quality testing, adaptive study design, drug discovery, crowd-sourcing, etc.

5. Conclusions

We provide a finite-time analysis for stratified sampling for Monte Carlo in the case of fixed strata with sub-Gaussian distributions. We report two bounds on the weighted MSE regret: (i) a distribution dependent bound of order $\text{poly}(\lambda_{\min}^{-1})\widetilde{O}(n^{-3/2})$ which is of interest when n is large compared to a measure of disparity $\lambda_{\min}^{-1}$ of the standard deviations (*stationary regime*), and (ii) a distribution free bound of order $\widetilde{O}(K^{1/3}n^{-4/3})$ which is of interest when n is small compared to $\lambda_{\min}^{-1}$ (*transitory regime*). We also link the weighted MSE loss to the total MSE loss of algorithm MC-UCB, that is the natural measure of performance for the problem. We provide $\mathrm{poly}(\lambda_{\min}^{-1})\widetilde{O}(n^{-3/2})$ problem-dependent and $\mathrm{poly}(K)\widetilde{O}(n^{-7/6})$ problem-independent bounds for the total MSE regret, as well. In case of symmetric arm distributions, the latter rate is improved to $\mathrm{poly}(K)\widetilde{O}(n^{-4/3})$. We give a distribution dependent bound of order $\mathrm{poly}(\lambda_{\min}^{-1})\widetilde{O}(n^{-3/2})$ and a distribution free bound of order $\widetilde{O}(K^{1/3}n^{-4/3})$ also on the pseudo-regret. The latter matches its minimax lower bound in terms of n up to a $\log n$ factor.

Possible directions for future work include: (i) making the MC-UCB algorithm anytime (i.e., not requiring the knowledge of n in advance) and (ii) deriving distribution-dependent lower bound for this problem determining the necessary dependence on $\lambda_{\min}$.

Acknowledgements This research was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270327 (CompLACS).

Appendices

These appendices contain the proofs of the theorems in the paper. Their organization is as follows.

- Appendix A contains the proofs of the (problem-dependent) Lemma 1) and the (problem-independent) Lemma 2) stating that the number of pulls of any arm is not too far from the optimal allocation for that arm on event ξ.

- Appendix B states some preliminary results which are useful in the regret bound proofs. It first gives (conditional) variance bound for sub-Gaussian random variables. Then it shows that ξ has high probability. It also contains the proof that for any $t \leq n$, $T_{k,t}$ is a stopping time, and applies Wald's identity to the samples from an arm. Next, it states bounds on some other technical quantities outside ξ that are used afterwards. Finally, it gives bounds on the parameters β_n and γ_n.[3]

- Appendix C contains the proofs of the (problem-dependent) Proposition 8 and the (problem-independent) Proposition 9 upper bounding $\widetilde{R}_n(\mathcal{A}_{\text{MC-UCB}})$ based on Lemma 1 and 2, respectively. These proofs are simpler than those in Appendix D and can serve as an introduction for the latter.

- Appendix D contains the proofs of the (problem-dependent) Proposition 3 and the (problem-independent) Proposition 4 upper bounding $R_n(\mathcal{A}_{\text{MC-UCB}})$ based on Lemma 1 and 2, respectively. These proofs are quite similar to the ones for bounding $\widetilde{R}_n(\mathcal{A}_{\text{MC-UCB}})$ in Appendix C. However, those have to be extended by additional technical steps, for example, using Wald's second identity for sums with random number of terms, to bound $R_n(\mathcal{A}_{\text{MC-UCB}})$ with a quantity reminding to $\widetilde{R}_n(\mathcal{A}_{\text{MC-UCB}})$.

- Appendix E provides the proofs of the three bounds on the cross product-terms. The first one holds when the arm distributions are symmetric: then the cross product-terms are exactly 0. The two other bounds, a problem-dependent and a problem-independent

[3] As for β, γ_n will be used for γ after this substituting $\delta = n^{-9/2}$.

one, concern the general sub-Gaussian case. These bounds rely on Lemma 1 and 2. Using these together with the results in Appendix D gives bounds on the total regret.

- Appendix F provides the proof of some general technical lemmata.

Appendix A. Proof of the Bounds on the Number of Pulls of the Arms

In this section, we prove Lemma 1 and 2. Recall that their statements hold on the event ξ. This event plays an important role in the proofs of the regret bounds; several statements will be proven on ξ. We transcribe the definition (13) of ξ into the following lemma when the number of samples $T_{k,t}$ are random.

Lemma 10 *For $k = 1, \ldots, K$ and $t = 2K, \ldots, n$, 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* (8) *and β be given by* (10)*. Then, on ξ, we have:*

$$|\hat{\sigma}_{k,t} - \sigma_k| \le \frac{2\beta}{\sqrt{T_{k,t}}}.$$

All statements in the proofs of this section are meant to hold on ξ.

A.1 Problem-Dependent Bound; Proof of Lemma 1

Proof of Lemma 1 The proof consists of the following three main steps.

Step 1. Properties of the algorithm. Recall the definition of the upper bound used in $\mathcal{A}_{\text{MC-UCB}}$ when $t > 2K$:

$$B_{q,t+1} = \frac{w_q}{T_{q,t}}\left(\hat{\sigma}_{q,t} + \frac{2\beta}{\sqrt{T_{q,t}}}\right), \qquad 1 \le q \le K.$$

From Lemma 10, we obtain the following upper and lower bounds for $B_{q,t+1}$ on ξ:

$$\frac{w_q \sigma_q}{T_{q,t}} \le B_{q,t+1} \le \frac{w_q}{T_{q,t}}\left(\sigma_q + \frac{4\beta}{\sqrt{T_{q,t}}}\right). \tag{20}$$

Note that as $n \ge 4K$, there exists an arm pulled after the initialization. Let k be such an arm and $t + 1 > 2K$ be the time step when k is pulled for the last time, that is, $T_{k,t} = T_{k,n} - 1 \ge 2$ and $T_{k,t+1} = T_{k,n}$. Since arm k is chosen at time $t + 1$, we have for any arm p

$$B_{p,t+1} \le B_{k,t+1}. \tag{21}$$

From (20) and the fact that $T_{k,t} = T_{k,n} - 1$, we obtain on ξ

$$B_{k,t+1} \le \frac{w_k}{T_{k,t}}\left(\sigma_k + \frac{4\beta}{\sqrt{T_{k,t}}}\right) = \frac{w_k}{T_{k,n} - 1}\left(\sigma_k + \frac{4\beta}{\sqrt{T_{k,n} - 1}}\right). \tag{22}$$

Using the lower bound in (20) and the fact that $T_{p,t} \le T_{p,n}$, we may lower bound $B_{p,t+1}$ on ξ as

$$B_{p,t+1} \ge \frac{w_p \sigma_p}{T_{p,t}} \ge \frac{w_p \sigma_p}{T_{p,n}}. \tag{23}$$

Combining (21), (22), and (23), we obtain on ξ

$$\frac{w_p\sigma_p}{T_{p,n}} \leq \frac{w_k}{T_{k,n}-1}\left(\sigma_k + \frac{4\beta}{\sqrt{T_{k,n}-1}}\right). \tag{24}$$

Note that at this point there is no dependency on t, and on ξ, (24) holds for any p and for any k such that $T_{k,n} > 2$.

Step 2. Lower bound on $T_{p,n}$. From the constraints $\sum_k (T_{k,n}-2) = n-2K$ and $\sum_k \lambda_k = 1$, we can deduce (by indirect proof) that there exists an arm k with $T_{k,n}-2 \geq \lambda_k(n-2K) > 0$, that is, $T_{k,n} > 2$. Thus k satisfies (24). Using (24), $T_{k,n}-1 > \lambda_k(n-2K)$, and $\lambda_k = w_k\sigma_k/\Sigma_w$ implies for any arm p

$$\frac{w_p\sigma_p}{T_{p,n}} < \frac{w_k}{n\lambda_k}\frac{1}{1-2K/n}\left(\sigma_k + \frac{4\beta}{\sqrt{n\lambda_k(1-2K/n)}}\right) \leq \frac{\Sigma_w}{n} + \frac{4K\Sigma_w}{n^2} + \frac{8\sqrt{2}\beta}{n^{3/2}\lambda_k^{3/2}},$$

because $n \geq 4K$. The previous inequality combined with the fact that $\lambda_k \geq \lambda_{\min}$ gives the first inequality (14) of the lemma

$$\frac{w_p\sigma_p}{T_{p,n}} \leq \frac{\Sigma_w}{n} + \frac{12\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}.$$

By rearranging it, we obtain the lower bound on $T_{p,n}$ in (15)

$$T_{p,n} \geq \frac{w_p\sigma_p}{\frac{\Sigma_w}{n} + \frac{12\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}} \geq T^*_{p,n} - 4\lambda_p\left(\frac{3\beta}{\Sigma_w\lambda_{\min}^{3/2}}\sqrt{n} + K\right), \tag{25}$$

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

Step 3. Upper bound on $T_{p,n}$. Using (25) and the fact that $\sum_k T_{k,n} = n$, we obtain

$$T_{p,n} = n - \sum_{k\neq p} T_{k,n} \leq \Big(n - \sum_{k\neq p} T^*_{k,n}\Big) + \sum_{k\neq p} 4\lambda_k\Big(\frac{3\beta}{\Sigma_w\lambda_{\min}^{3/2}}\sqrt{n} + K\Big).$$

Since $\sum_{k\neq p}\lambda_k \leq 1$ and $\sum_k T^*_{k,n} = n$, we deduce

$$T_{p,n} \leq T^*_{p,n} + 4\left(\frac{3\beta}{\Sigma_w\lambda_{\min}^{3/2}}\sqrt{n} + K\right). \tag{26}$$

The lemma follows by combining the lower and upper bounds in (25) and (26). ■

A.2 Problem-Independent Bound; Proof of Lemma 2

Proof of Lemma 2

Step 1. Lower bound of order $\Omega(n^{2/3})$. Recall the definition of the upper bound $B_{q,t+1}$ used in $\mathcal{A}_{\text{MC-UCB}}$ when $t > 2K$:

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

Using Lemma 10 it follows that on ξ, for any q such that $T_{q,t} \geq 2$,

$$\frac{w_q \sigma_q}{T_{q,t}} \leq B_{q,t+1} \leq \frac{w_q}{T_{q,t}}\left(\sigma_q + \frac{4\beta}{\sqrt{T_{q,t}}}\right). \tag{27}$$

Let k be the index of an arm that is such that $T_{k,n} - 2 \geq w_k(n-2K)$. Such an arm always exists for any possible allocation strategy, as $n - 2K = \sum_q (T_{q,n} - 2)$ and $\sum_q w_q = 1$. This implies $T_{k,n} \geq 3$ as $n \geq 4K$, thus arm k is pulled after the initialization. Let $t+1 \leq n$ be the last time at which it was pulled, that is, $T_{k,t} = T_{k,n} - 1$ and $T_{k,t+1} = T_{k,n}$. From (27) and the fact that $T_{k,t} > w_k(n-2K)$ and $T_{k,t} \geq 2$, we obtain on ξ

$$B_{k,t+1} \leq \frac{w_k}{T_{k,t}}\left(\sigma_k + \frac{4\beta}{\sqrt{T_{k,t}}}\right) < \frac{\max_p \sigma_p + \sqrt{8}\beta}{n - 2K}. \tag{28}$$

Since at time $t+1$ the arm k has been pulled, then for any arm q, we have

$$B_{q,t+1} \leq B_{k,t+1}. \tag{29}$$

From the definition of $B_{q,t+1}$, and also using the fact that $T_{q,t} \leq T_{q,n}$, we deduce on ξ that

$$B_{q,t+1} \geq \frac{2\beta w_q}{T_{q,t}^{3/2}} \geq \frac{2\beta w_q}{T_{q,n}^{3/2}}. \tag{30}$$

Combining (28)–(30), we obtain on ξ

$$\frac{2\beta w_q}{T_{q,n}^{3/2}} < \frac{\max_p \sigma_p + \sqrt{8}\beta}{n - 2K} = \frac{\bar{\Sigma} + \sqrt{8}\beta}{n - 2K}.$$

Finally, this implies on ξ that for any q,

$$T_{q,n} \geq \left(\frac{2\beta w_q (n-2K)}{\bar{\Sigma} + \sqrt{8}\beta}\right)^{2/3} = \left(\frac{2 - 4K/n}{\bar{\Sigma}/\beta + \sqrt{8}} w_q n\right)^{2/3} \geq \frac{(w_q n)^{2/3}}{(\bar{\Sigma}/\beta + \sqrt{8})^{2/3}} = \frac{(w_q n)^{2/3}}{\gamma^2}$$

recalling $\gamma = (\bar{\Sigma}/\beta + \sqrt{8})^{1/3}$, which proves (16).

Step 2. Properties of the algorithm. We follow a similar analysis to Step 1 of the proof of Lemma 1. Note that as $n \geq 4K$, there exists an arm pulled after the initialization. Let q be any such arm and $t+1 > 2K$ be the time step when q is pulled for the last time, that is, $T_{q,t} = T_{q,n} - 1 \geq 2$ and $T_{q,t+1} = T_{q,n}$. Since arm q is chosen at time $t+1$, we have for any arm p

$$B_{p,t+1} \leq B_{q,t+1}. \tag{31}$$

From (27) and $T_{q,t} = T_{q,n} - 1$, we obtain on ξ

$$B_{q,t+1} \leq \frac{w_q}{T_{q,t}}\left(\sigma_q + \frac{4\beta}{\sqrt{T_{q,t}}}\right) = \frac{w_q}{T_{q,n}-1}\left(\sigma_q + \frac{4\beta}{\sqrt{T_{q,n}-1}}\right). \tag{32}$$

Furthermore, since $T_{p,t} \leq T_{p,n}$ and $T_{p,t} \geq 2$ (as $t \geq 2K$), then on ξ

$$B_{p,t+1} \geq \frac{w_p\sigma_p}{T_{p,t}} \geq \frac{w_p\sigma_p}{T_{p,n}}. \tag{33}$$

Combining (31)–(33), we obtain on ξ

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

Note that this inequality holds on ξ for any p and for any q such that $T_{q,n} \geq 3$. Summing over all such q on both sides, we obtain on ξ for any arm p

$$\frac{w_p\sigma_p}{T_{p,n}}\sum_{q:T_{q,n}\geq 3}(T_{q,n}-1) \leq \sum_{q:T_{q,n}\geq 3} w_q\left(\sigma_q + \frac{4\beta}{\sqrt{T_{q,n}-1}}\right).$$

This implies

$$\frac{w_p\sigma_p}{T_{p,n}}(n-2K) \leq \sum_{q=1}^{K} w_q\left(\sigma_q + \frac{4\beta}{\sqrt{T_{q,n}-1}}\right), \tag{34}$$

because $\sum_{q:T_{q,n}\geq 3}(T_{q,n}-1) = n - K - \sum_{q:T_{q,n}\leq 2}(T_{q,n}-1) \geq n-K-K = n-2K$.

Step 3. Lower bound. Plugging (16) into (34),

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

on ξ, since $T_{q,n}-1 \geq \frac{T_{q,n}}{2}$ (as $T_{q,n} \geq 2$) and because $\sum_q w_q^{2/3} \leq K^{1/3}$ by Jensen's inequality. Finally as $n \geq 4K$, we obtain on ξ the first inequality (17) of the lemma

$$\frac{w_p\sigma_p}{T_{p,n}} \leq \frac{\Sigma_w}{n} + \frac{12K^{1/3}\beta\gamma}{n^{4/3}} + \frac{4K\Sigma_w}{n^2}.$$

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

$$T_{p,n} \geq \frac{w_p\sigma_p}{\frac{\Sigma_w}{n} + 12K^{1/3}\beta\gamma n^{-4/3} + \frac{4K\Sigma_w}{n^2}} \geq T^*_{p,n} - 4\lambda_p\left(\frac{3K^{1/3}\beta\gamma}{\Sigma_w}n^{2/3} + K\right),$$

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

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

$$T_{p,n} \leq n - \sum_{q \neq p} T^*_{q,n} + \sum_{q \neq p} 4\lambda_q \left(\frac{3K^{1/3}\beta\gamma}{\Sigma_w} n^{2/3} + K \right) \leq T^*_{p,n} + 4 \left(\frac{3K^{1/3}\beta\gamma}{\Sigma_w} n^{2/3} + K \right),$$

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

Appendix B. Main Tools for the Bounds on the Regrets

In this section, we first give a high probability uniform upper bound on the estimation errors of the unbiased empirical standard deviations for sub-Gaussian random variables, then describe other technical tools, properties, and inequalities. Several of these use the following simple lemma giving (conditional) variance bound for sub-Gaussian random variables proven in Appendix F:

Lemma 11 *Let A be an event with $\mathbb{P}(A) \leq \delta$. Let X have a distribution with $\mu \overset{\text{def}}{=} \mathbb{E}X$ satisfying* (9) *of Assumption 1 with $c_1 > 0$, $c_2 \geq \delta$, and any $\epsilon > 0$. Then*

$$\mathbb{E}\left[|X - \mu|^2 \mathbb{I}\{A\} \right] \leq \delta c_1 \log(ec_2/\delta).$$

Particularly, the case $\mathbb{P}(A) = \delta = 1$ gives $\operatorname{Var} X \leq c_1 \log(ec_2)$ *if* $c_2 \geq 1$.

B.1 High Probability Uniform Upper Bound on the Variance Estimation Errors

In this subsection, let $n \geq 2$, $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ, variance σ^2, and unbiased empirical variances

$$\hat{s}_t^2 = \frac{1}{t-1} \sum_{i=1}^{t} \Big(X_i - \frac{1}{t} \sum_{t'=1}^{t} X_{t'} \Big)^2 \tag{35}$$

corresponding to the first t variables, and also $\hat{s}_t = \sqrt{\hat{s}_t^2}$ $(2 \leq t \leq n)$.

The upper confidence bounds $B_{k,t}$ used in the MC-UCB algorithm is motivated by the following theorem of Maurer and Pontil (2009) (see also the paper of Audibert et al., 2009, for a variant), that gives a high probability bound on the estimation error of $\hat{s}_t$:

Theorem 12 (Theorem 10 of Maurer and Pontil, 2009) *If $\forall t \leq n$, $X_t \in [a, a+c]$, then for $0 < \delta \leq 2$, with probability at least $1 - \delta$*

$$|\hat{s}_n - \sigma| \leq c\sqrt{\frac{2\log(2/\delta)}{n-1}}.$$

Using the union bound and $t/(t-1) \leq 2$ for $t \geq 2$ this implies the following uniform bound:

Corollary 13 *If $\forall t \leq n$, $X_t \in [a, a+c]$, then for $0 < \delta \leq 2$, the event*

$$\bigcap_{2 \leq t \leq n} \left\{ |\hat{s}_t - \sigma| \leq 2c\sqrt{\frac{\log(2/\delta)}{t}} \right\}.$$

has probability at least $1 - n\delta$.

We extend this result to sub-Gaussian random variables:

Proposition 14 *Let the distribution of X_t's satisfy (9) of Assumption 1 with $c_1 > 0$, $c_2 \geq 1$, and any $\epsilon > 0$. Define the following event for any $0 < \delta < 1/e$*

$$\xi_n(\delta) = \bigcap_{2 \leq t \leq n} \left\{ |\hat{s}_t - \sigma| \leq \frac{2\beta}{\sqrt{t}} \right\},$$

where β is given by (10). Then $\mathbb{P}(\xi_n(\delta)) \geq (1 - n\delta)^2$.

Proof of Proposition 14 *Step 1. Truncating sub-Gaussian variables.* Let the conditional variance of X_t be $\tilde{\sigma}^2 \overset{\text{def}}{=} \text{Var}[X_t | (X_t - \mu)^2 \leq c_1 \log(c_2/\delta)]$. We characterize $\tilde{\sigma}$ by the following lemma (proven in Appendix F):

Lemma 15 *Let X have a distribution with $\mu \overset{\text{def}}{=} \mathbb{E}X$ and $\sigma^2 \overset{\text{def}}{=} \text{Var}\, X$ satisfying (9) of Assumption 1 with $c_1 > 0$, $c_2 \geq 1$, and any $\epsilon > 0$. Let $0 < \delta < 1/e$, $A \overset{\text{def}}{=} \{|X - \mu|^2 \leq c_1 \log(c_2/\delta)\}$, and $\tilde{\sigma}^2 \overset{\text{def}}{=} \text{Var}[X|A]$. Then $\mathbb{P}(A^C) \leq \delta$ and*

$$0 \leq \sigma - \tilde{\sigma} \leq \frac{\sqrt{c_1 \delta \log(ec_2/\delta)}}{1 - \delta}.$$

Step 2. Application of tail inequalities. Define the event

$$\xi_1 = \xi_{1,n}(\delta) = \bigcap_{1 \leq t \leq n} \left\{ |X_t - \mu|^2 \leq c_1 \log(c_2/\delta) \right\}.$$

We have that $\mathbb{P}(\xi_1^C) \leq n\delta$ using the union bound and (9). Given ξ_1, $(X_t)_{1 \leq t \leq n}$ are n i.i.d. bounded random variables with common conditional variance $\tilde{\sigma}^2$.

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

$$\xi_2 = \bigcap_{2 \leq t \leq n} \left\{ |\hat{s}_t - \tilde{\sigma}| \leq 4\sqrt{c_1 \log(c_2/\delta) \frac{\log(2/\delta)}{t}} \right\}.$$

Using Corollary 13, we deduce that $\mathbb{P}(\xi_2 | \xi_1) \geq 1 - n\delta$, and thus

$$\mathbb{P}(\xi_1 \cap \xi_2) = \mathbb{P}(\xi_2 | \xi_1)\mathbb{P}(\xi_1) \geq (1 - n\delta)^2.$$

Moreover, from Lemma 15, we have $0 \leq \sigma - \tilde{\sigma} \leq \frac{\sqrt{c_1 \delta \log(ec_2/\delta)}}{1-\delta}$, and thus on ξ_2, for all $2 \leq t \leq n$:

$$|\hat{s}_t - \sigma| \leq |\hat{s}_t - \tilde{\sigma}| + |\tilde{\sigma} - \sigma| \leq 4\sqrt{c_1 \log(c_2/\delta) \frac{\log(2/\delta)}{t}} + \frac{\sqrt{c_1 \delta \log(ec_2/\delta)}}{1 - \delta} \leq \frac{2\beta}{\sqrt{t}},$$

implying $\xi_2 \subseteq \xi_n(\delta)$. From this, we deduce

$$\mathbb{P}(\xi_n(\delta)) \geq \mathbb{P}(\xi_2) \geq \mathbb{P}(\xi_1 \cap \xi_2) \geq (1 - n\delta)^2$$

proving the proposition. ■

Corollary 16 *Let $n \geq 2$. Let Assumption 1 hold with $c_1 > 0$, $c_2 \geq 1$, and any $\epsilon > 0$. For any $0 < \delta < 1/e$ and for event ξ defined by* (13)*, $\mathbb{P}(\xi) \geq (1 - n\delta)^{2K} \geq 1 - 2nK\delta$.*

Proof of Corollary 16 Since for each $1 \leq k \leq K$, Proposition 14 implies that the probability of

$$\bigcap_{2 \leq t \leq n} \left\{ |\hat{s}_{k,t} - \sigma_k| \leq \frac{2\beta}{\sqrt{t}} \right\}$$

is at least $(1 - n\delta)^2$, the intersection of these independent events, ξ, has probability at least $(1 - n\delta)^{2K}$. The last inequality comes from the convexity of $(1 - x)^{2K}$. ■

B.2 $T_{k,t}$ is Stopping Time, Wald's Identity for the Variance of the Sum of $T_{k,t}$ Centered Samples of One Arm

For a given k, let $(\mathcal{F}_t^{(k)})_{t \leq n}$ be the filtration associated to the process $(X_{k,t})_{t \leq n}$, and $\mathcal{E}_{-k} = \mathcal{E}_{-k,n}$ be the σ-algebra generated by $(X_{k',t'})_{t' \leq n, k' \neq k}$ ("environment"). Define the filtration $(\mathcal{G}_t^{(k)})_{t \leq n}$ by

$$\mathcal{G}_t^{(k)} = \mathcal{G}_t^{(k,n)} \overset{\text{def}}{=} \sigma(\mathcal{F}_t^{(k)}, \mathcal{E}_{-k}).$$

$T_{k,t}$ is a stopping time. We prove the following proposition.

Proposition 17 *For each $1 \leq n' \leq n$, $T_{k,n'}$ is a stopping time w.r.t. $(\mathcal{G}_t^{(k)})_{t \leq n}$.*

Proof We prove the statement for fixed budget n by induction for $n' = 1, \ldots, n$.

For $n' \leq 2K$ (initialization), $T_{k,n'}$ is deterministic, so for any t, $\{T_{k,n'} \leq t\}$ is either the empty set or the whole probability space (and is thus measurable according to $\mathcal{G}_t^{(k)}$).

Let us now assume that for a given time step $2K \leq n' < n$, and for any t, $\{T_{k,n'} \leq t\}$ is $\mathcal{G}_t^{(k)}$-measurable. We consider now time step $n' + 1$. Note first that for $t = 0$, $\{T_{k,n'+1} \leq t\} = \{T_{k,n'+1} \leq 0\}$ is the empty set and is thus $\mathcal{G}_t^{(k)}$-measurable. If $t > 0$, then

$$\{T_{k,n'+1} \leq t\} = \big(\{T_{k,n'} = t\} \cap \{k_{n'+1} \neq k\}\big) \cup \{T_{k,n'} \leq t - 1\}. \tag{36}$$

By induction assumption, $\{T_{k,n'} = t\}$ and $\{T_{k,n'} \leq t - 1\}$ are $\mathcal{G}_t^{(k)}$-measurable (since *for any t'*, $\{T_{k,n'} \leq t'\}$ is $\mathcal{G}_{t'}^{(k)}$-measurable). On $\{T_{k,n'} = t\}$, $k_{n'+1}$ is also $\mathcal{G}_t^{(k)}$-measurable since it is determined only by the values of the upper bounds $\{B_{q,n'+1}\}_{1 \leq q \leq K}$ (which depend only on $\{X_{k',t'}\}_{t' \leq n, k' \neq k}$ and on $(X_{k,1}, \ldots, X_{k,t})$). Hence, $\{T_{k,n'} = t\} \cap \{k_{n'+1} \neq k\}$ is $\mathcal{G}_t^{(k)}$-measurable, and thus using (36), we have that $\{T_{k,n'+1} \leq t\}$ is $\mathcal{G}_t^{(k)}$-measurable, as well.

We have thus proved by induction that $T_{k,n'}$ is a stopping time w.r.t. the filtration $(\mathcal{G}_t^{(k)})_{t\leq n}$. ■

Wald's second identity for the variance. We also need to express the variance of the sum of random number of centered terms when this random number is a stopping time. Thus, we recall the following theorem from Athreya and Lahiri (2006) (this variant is quoted from Lemma 10 of Antos et al. (2010))

Proposition 18 (Theorem 13.2.14 of Athreya and Lahiri (2006)) *Let $(\mathcal{F}_t)_{t=1,\dots,n}$ be a filtration and $(X_t)_{t=1,\dots,n}$ be an $\mathcal{F}_t$ adapted sequence of i.i.d. random variables with finite expectation μ and variance σ^2. Assume that $\mathcal{F}_t$ and $\sigma(\{X_s : s \geq t+1\})$ are independent for any $t \leq n$, and let $T(\leq n)$ be a stopping time w.r.t. $\mathcal{F}_t$. Then*

$$\mathbb{E}\bigg[\Big(\sum_{t=1}^{T}(X_t-\mu)\Big)^2\bigg] = \mathbb{E}[T]\sigma^2.$$

Application to arm k and samples $(X_{k,t})_{t\leq n}$.

Corollary 19 *For any $1 \leq k \leq K$ and $n' \leq n$,*

$$\mathbb{E}\left[\Big(\sum_{t=1}^{T_{k,n'}}(X_{k,t}-\mu_k)\Big)^2\right] = \mathbb{E}[T_{k,n'}]\sigma_k^2.$$

Proof Proposition 17, the fact that $G_t^{(k)}$ and $\sigma(\{X_{k,s} : s \geq t+1\})$ are independent for any $t \leq n$, and $T_{k,n'} \leq n$ guarantee that we can apply Proposition 18 with filtration $(\mathcal{G}_t^{(k)})_{t\leq n}$, $(X_{k,t})_{t\leq n}$, and $T_{k,n'}$ leading to the equality. ■

B.3 Other Technical Inequalities

Now we state and prove some further technical inequalities.

Bounds on the loss and the variance of the sum of the centered samples of one arm on event ξ^C.

Lemma 20 *Let $n \geq 2$ and $0 < \delta < 1/e$. Let Assumption 1 hold with $c_2 \geq \max(1, 2nK\delta)$. Then for each arm k,*

$$\mathbb{E}\big[|\hat{\mu}_{k,n}-\mu_k|^2\mathbb{I}\{\xi^C\}\big] \leq Kn^2\delta C_\xi(\delta) \qquad \textit{and}$$

$$\mathbb{E}\left[\Big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\Big)^2\mathbb{I}\{\xi^C\}\right] \leq 2Kn^3\delta C_\xi(\delta),$$

where $C_\xi(\delta) = C_{\xi,n}(\delta) \stackrel{\text{def}}{=} c_1 \log(ec_2/2nK\delta)$. *Consequently, for every arms* k *and* q,

$$\big|\mathbb{E}\big[(\hat\mu_{k,n}-\mu_k)(\hat\mu_{q,n}-\mu_q)\mathbb{I}\{\xi^C\}\big]\big| \le Kn^2\delta C_\xi(\delta) \qquad \textit{and}$$

$$\left|\mathbb{E}\left[\Big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\Big)\Big(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\Big)\mathbb{I}\{\xi^C\}\right]\right| \le 2Kn^3\delta C_\xi(\delta).$$

Proof of Lemma 20 $c_2 \ge 1$ and Corollary 16 imply $\mathbb{P}(\xi^C) \le 2nK\delta$. Due to this, $c_2 \ge 2nK\delta$, and Assumption 1, for any $1 \le k \le K$ and $1 \le t \le n$, Lemma 11 implies

$$\mathbb{E}\left[(X_{k,t}-\mu_k)^2\mathbb{I}\{\xi^C\}\right] \le 2nK\delta c_1 \log(ec_2/2nK\delta) = 2Kn\delta C_\xi(\delta).$$

The first claim follows from the fact that

$$\begin{aligned}\mathbb{E}\big[(\hat\mu_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}\big] &\le \mathbb{E}\left[\frac{\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)^2}{T_{k,n}}\mathbb{I}\{\xi^C\}\right]\\ &\le \sum_{t=1}^{n}\frac{\mathbb{E}\big[(X_{k,t}-\mu_k)^2\mathbb{I}\{\xi^C\}\big]}{2} \le Kn^2\delta C_\xi(\delta).\end{aligned}$$

The second claim follows from the fact that

$$\Big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\Big)^2 \le \Big(\sum_{t=1}^{n}|X_{k,t}-\mu_k|\Big)^2 \le n\sum_{t=1}^{n}(X_{k,t}-\mu_k)^2,$$

and so

$$\mathbb{E}\bigg[\Big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\Big)^2\mathbb{I}\{\xi^C\}\bigg] \le n\sum_{t=1}^{n}\mathbb{E}\big[(X_{k,t}-\mu_k)^2\mathbb{I}\{\xi^C\}\big] \le 2Kn^3\delta C_\xi(\delta).$$

The third claim follows from the first one by Cauchy-Schwarzs inequality

$$\big|\mathbb{E}\big[(\hat\mu_{k,n}-\mu_k)(\hat\mu_{q,n}-\mu_q)\mathbb{I}\{\xi^C\}\big]\big| \le \sqrt{\mathbb{E}[(\hat\mu_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}]}\sqrt{\mathbb{E}[(\hat\mu_{q,n}-\mu_q)^2\mathbb{I}\{\xi^C\}]},$$

and the fourth one follows from the second one, analogously. ■

We get the following corollary by substituting $\delta = n^{-9/2}$:

Corollary 21 *Let* $n \ge K \ge 2$. *Let Assumption 1 hold with* $c_2 \ge 1$. *Then for each arm* k,

$$\mathbb{E}\big[|\hat\mu_{k,n}-\mu_k|^2\mathbb{I}\{\xi^C\}\big] \le \frac{KC_\xi}{n^{5/2}} \qquad \textit{and}$$

$$\mathbb{E}\left[\Big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\Big)^2\mathbb{I}\{\xi^C\}\right] \le \frac{2KC_\xi}{n^{3/2}}.$$

where $C_\xi = C_\xi(n^{-9/2}) = c_1 \log(ec_2 n^{7/2}/2K)$ *as in* (19). *Consequently, for every arms* k *and* q,

$$\big|\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{q,n} - \mu_q)\mathbb{I}\{\xi^C\}\big]\big| \le \frac{KC_\xi}{n^{5/2}} \qquad \textit{and}$$

$$\left|\mathbb{E}\left[\Big(\sum_{t=1}^{T_{k,n}}(X_{k,t} - \mu_k)\Big)\Big(\sum_{t=1}^{T_{q,n}}(X_{q,t} - \mu_q)\Big)\mathbb{I}\{\xi^C\}\right]\right| \le \frac{2KC_\xi}{n^{3/2}}.$$

Upper and lower bound on β_n *of* (11) *for* $\delta = n^{-9/2}$. Using $n \ge 4K \ge 8$, $c_2 \ge 1$, and monotonicity in n we have

$$\begin{aligned}
\beta_n &= \sqrt{c_1 \log(c_2^2 n^9)\log(4n^9)} + \frac{\sqrt{c_1 \log(ec_2 n^{4.5})}}{2(1-n^{-4.5})n^{7/4}} \\
&\le \sqrt{c_1}\frac{\log(c_2^2 n^9) + \log(4n^9)}{2} + \frac{\sqrt{c_1 \log(ec_2 8^{4.5})}}{2(1-8^{-4.5})8^{7/4}} \\
&\le \sqrt{c_1}\left(9\log n + \log(4c_2^2)/2 + \frac{\log(e^2c_2^2 8^9)}{2^{5/4}(8^2 - 8^{-2.5})\sqrt{\log(e8^{4.5})}}\right).
\end{aligned}$$

Using

$$\log(e^2c_2^2 8^9) \le \log(e^2c_2^2(c_2+1)^{27}) \le 29\log(c_2+1) + 2\log(c_2+1)/\log 2 \le 32\log(c_2+1)$$

and $4c_2^2 \le (c_2+1)^3$

$$\beta_n \le \sqrt{c_1}\left(9\log n + 1.5\log(c_2+1) + \frac{32\log(c_2+1)}{489}\right) \le \sqrt{c_1}(9\log n + 1.6\log(c_2+1)) = C_\beta$$

recalling (18). On the other hand, keeping only the first term of β_n

$$\beta_n \ge \sqrt{c_1\log(c_2^2n^9)\log(4n^9)} \ge \sqrt{c_1\log(8^9c_2^2)29\log 2} \ge \sqrt{58c_1\log 2\log(ec_2)} \ge \sqrt{40c_1\log(ec_2)}.$$

Upper bound on γ_n *of Lemma 2 when* $\delta = n^{-9/2}$. If Assumption 1 is satisfied with $c_2 \ge 1$ then Lemma 11 implies $\sigma_k^2 \le c_1\log(ec_2)$ for any $1 \le k \le K$, thus recalling $\bar{\Sigma} = \max_p \sigma_p$ we have $\Sigma_w \le \bar{\Sigma} \le \sqrt{c_1\log(ec_2)}$. For $\delta = n^{-9/2}$, the lower bound above on β_n leads to $\bar{\Sigma}/\beta_n \le 1/\sqrt{40}$ and

$$\gamma_n = (\bar{\Sigma}/\beta_n + \sqrt{8})^{1/3} \le (1/\sqrt{40} + \sqrt{8})^{1/3} < 1.5.$$

Appendix C. Proof of Proposition 8 and 9

In this section, we use Lemmata 1 and 2 to prove Proposition 8 and 9, respectively.

C.1 Proof of Proposition 8

Proof of Proposition 8 By definition, the pseudo-loss of the algorithm is

$$\begin{aligned}\widetilde{L}_n(\mathcal{A}_{\text{MC-UCB}}) &= \sum_{k=1}^K w_k^2\sigma_k^2 \mathbb{E}\Big[\frac{\mathbb{I}\{\xi\}}{T_{k,n}}\Big] + \sum_{k=1}^K w_k^2\sigma_k^2 \mathbb{E}\Big[\frac{\mathbb{I}\{\xi^C\}}{T_{k,n}}\Big] \\ &\leq \sum_{k=1}^K \frac{w_k^2\sigma_k^2\mathbb{P}(\xi)}{\inf_{\omega\in\xi} T_{k,n}(\omega)} + \sum_{k=1}^K w_k^2\sigma_k^2\frac{\mathbb{P}(\xi^C)}{2}, \qquad (37)\end{aligned}$$

because $T_{k,n} \geq 2$ by the definition of $\mathcal{A}_{\text{MC-UCB}}$. Recalling (14) from Lemma 1 that upper bounds $w_k\sigma_k/T_{k,n}$ on ξ, we obtain

$$\sum_{k=1}^K \frac{w_k^2\sigma_k^2\mathbb{P}(\xi)}{\inf_\xi T_{k,n}} \leq \sum_{k=1}^K w_k\sigma_k\Big(\frac{\Sigma_w}{n} + \frac{12\beta_n}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}\Big) = \frac{\Sigma_w^2}{n} + \frac{12\Sigma_w\beta_n}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w^2}{n^2}$$

using $\sum_k w_k\sigma_k = \Sigma_w$. Finally, using (37) and the previous inequality and recalling $\mathbb{P}(\xi^C) \leq 2nK\delta$ from Corollary 16, $\delta = n^{-9/2}$, and $\beta_n \leq C_\beta$ from Appendix B.3, we have

$$\begin{aligned}\widetilde{R}_n(\mathcal{A}_{\text{MC-UCB}}) &= \widetilde{L}_n(\mathcal{A}_{\text{MC-UCB}}) - \frac{\Sigma_w^2}{n} \\ &\leq \frac{12\Sigma_w\beta_n}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w^2}{n^2} + nK\delta\sum_{k=1}^K w_k^2\sigma_k^2 \\ &\leq \frac{12\Sigma_w C_\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w^2}{n^2} + \frac{K\Sigma_w^2}{n^{7/2}} \\ &\leq \frac{12\Sigma_w C_\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{(4K+\sqrt{2}/16)\Sigma_w^2}{n^2},\end{aligned}$$

that concludes the proof. ■

C.2 Proof of Proposition 9

Proof of Proposition 9 We decompose and bound $\widetilde{L}_n(\mathcal{A}_{\text{MC-UCB}})$ on ξ and ξ^C again as in (37). Recalling (17) from Lemma 2 that upper bounds $w_k\sigma_k/T_{k,n}$ on ξ, we obtain

$$\sum_{k=1}^K \frac{w_k^2\sigma_k^2\mathbb{P}(\xi)}{\inf_\xi T_{k,n}} \leq \sum_{k=1}^K w_k\sigma_k\Big(\frac{\Sigma_w}{n} + \frac{12K^{1/3}\beta_n\gamma_n}{n^{4/3}} + \frac{4K\Sigma_w}{n^2}\Big) = \frac{\Sigma_w^2}{n} + \frac{12K^{1/3}\Sigma_w\beta_n\gamma_n}{n^{4/3}} + \frac{4K\Sigma_w^2}{n^2}$$

using $\sum_k w_k \sigma_k = \Sigma_w$. Finally, using (37) and the previous inequality and recalling $\mathbb{P}(\xi^C) \leq 2nK\delta$ from Corollary 16, $\delta = n^{-9/2}$, $\beta_n \leq C_\beta$, and $\gamma_n < 1.5$ from Appendix B.3, we have

$$\begin{aligned}\widetilde{R}_n(\mathcal{A}_{\text{MC-UCB}}) &= \widetilde{L}_n(\mathcal{A}_{\text{MC-UCB}}) - \frac{\Sigma_w^2}{n}\\ &\leq \frac{12K^{1/3}\Sigma_w\beta_n\gamma_n}{n^{4/3}} + \frac{4K\Sigma_w^2}{n^2} + nK\delta\sum_{k=1}^K w_k^2\sigma_k^2\\ &\leq \frac{18K^{1/3}\Sigma_w C_\beta}{n^{4/3}} + \frac{4K\Sigma_w^2}{n^2} + \frac{K\Sigma_w^2}{n^{7/2}}\\ &\leq \frac{18K^{1/3}\Sigma_w C_\beta}{n^{4/3}} + \frac{(4K+\sqrt{2}/16)\Sigma_w^2}{n^2},\end{aligned}$$

that concludes the proof. ■

Appendix D. Bounds on $R_n(\mathcal{A}_{\textbf{MC-UCB}})$

This section contains the proofs of the regret bounds for $\mathcal{A}_{\text{MC-UCB}}$.

D.1 Problem-Dependent Bound

Proof of Proposition 3 By definition, we have

$$L_n(\mathcal{A}_{\text{MC-UCB}}) = \sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\Big] + \sum_{k=1}^K w_k^2\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}\Big]. \tag{38}$$

Using the definition of $\hat{\mu}_{k,n}$, we have

$$(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\} \leq \frac{\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)^2}{\inf_{\omega\in\xi} T_{k,n}^2(\omega)}\mathbb{I}\{\xi\} \leq \frac{\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)^2}{\inf_\xi T_{k,n}^2}.$$

Taking expectation and using Corollary 19

$$\mathbb{E}\big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\big] \leq \frac{\mathbb{E}\Big[\big(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\big)^2\Big]}{\inf_\xi T_{k,n}^2} = \frac{\mathbb{E}[T_{k,n}]\sigma_k^2}{\inf_\xi T_{k,n}^2},$$

so we bound the first sum of (38) as

$$\sum_{k=1}^K w_k^2\mathbb{E}\big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\big] \leq \sum_{k=1}^K w_k^2\frac{\sigma_k^2\mathbb{E}[T_{k,n}]}{\inf_\xi T_{k,n}^2}. \tag{39}$$

Recalling (14) from Lemma 1 that upper bounds $w_k\sigma_k/T_{k,n}$ on ξ, we obtain

$$\sum_{k=1}^K w_k^2\frac{\sigma_k^2\mathbb{E}[T_{k,n}]}{\inf_\xi T_{k,n}^2} \leq \sum_{k=1}^K\Big(\frac{\Sigma_w}{n} + \frac{12\beta_n}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}\Big)^2\mathbb{E}[T_{k,n}]. \tag{40}$$

Since $\sum_k T_{k,n} = n$, we have $\sum_k \mathbb{E}[T_{k,n}] = n$, (40) can be rewritten as

$$\begin{aligned}
\sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2 \mathbb{E}[T_{k,n}]}{\inf_\xi T_{k,n}^2} &\le \Big(\frac{\Sigma_w}{n} + \frac{12\beta_n}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{4K\Sigma_w}{n^2}\Big)^2 n \\
&\le \Big(\frac{\Sigma_w^2}{n^2} + \frac{24\Sigma_w\beta_n}{n^{5/2}\lambda_{\min}^{3/2}} + \frac{8K\Sigma_w^2}{n^3} + \frac{288\beta_n^2}{n^3\lambda_{\min}^3} + \frac{32K^2\Sigma_w^2}{n^4}\Big) n \\
&\le \frac{\Sigma_w^2}{n} + \frac{24\Sigma_w\beta_n}{n^{3/2}\lambda_{\min}^{3/2}} + 16\frac{K\Sigma_w^2}{n^2} + \frac{288\beta_n^2}{n^2\lambda_{\min}^3}.
\end{aligned}$$

Finally, using (38), (39), and the previous inequality and recalling $\delta = n^{-9/2}$, Corollary 21, and $\beta_n \le C_\beta$ from Appendix B.3 we have

$$\begin{aligned}
R_n(\mathcal{A}_{\text{MC-UCB}}) &= L_n(\mathcal{A}_{\text{MC-UCB}}) - \frac{\Sigma_w^2}{n} \\
&\le \frac{24\Sigma_w\beta_n}{n^{3/2}\lambda_{\min}^{3/2}} + 16\frac{K\Sigma_w^2}{n^2} + \frac{288\beta_n^2}{n^2\lambda_{\min}^3} + \frac{KC_\xi}{n^{5/2}} \\
&\le \frac{24\Sigma_w C_\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{288C_\beta^2}{n^2\lambda_{\min}^3} + 16\frac{K\Sigma_w^2}{n^2} + \frac{\sqrt{K}C_\xi}{2n^2} \\
&\le \frac{24\Sigma_w C_\beta}{n^{3/2}\lambda_{\min}^{3/2}} + \frac{288C_\beta^2}{n^2\lambda_{\min}^3} + \frac{\sqrt{K}C_\xi + 32K\Sigma_w^2}{2n^2}.
\end{aligned}$$

This concludes the proof. ■

D.2 Problem-Independent Bound

Proof of Proposition 4 Again, we decompose $L_n(\mathcal{A}_{\text{MC-UCB}})$ on ξ and ξ^C as in (38), and bound it on ξ as in (39). Recalling (17) from Lemma 2 that upper bounds $w_k\sigma_k/T_{k,n}$ on ξ, we obtain

$$\sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2 \mathbb{E}[T_{k,n}]}{\inf_\xi T_{k,n}^2} \le \sum_{k=1}^{K}\Big(\frac{\Sigma_w}{n} + \frac{12K^{1/3}\beta_n\gamma_n}{n^{4/3}} + \frac{4K\Sigma_w}{n^2}\Big)^2 \mathbb{E}[T_{k,n}]. \tag{41}$$

Since $\sum_k T_{k,n} = n$, we have $\sum_k \mathbb{E}[T_{k,n}] = n$, (41) can be rewritten as

$$\begin{aligned}
\sum_{k=1}^{K} w_k^2 \frac{\sigma_k^2 \mathbb{E}[T_{k,n}]}{\inf_\xi T_{k,n}^2} &\le \Big(\frac{\Sigma_w}{n} + \frac{12K^{1/3}\beta_n\gamma_n}{n^{4/3}} + \frac{4K\Sigma_w}{n^2}\Big)^2 n \\
&\le \Big(\frac{\Sigma_w^2}{n^2} + \frac{24K^{1/3}\Sigma_w}{n^{7/3}}\beta_n\gamma_n + \frac{8K\Sigma_w^2}{n^3} + \frac{288K^{2/3}}{n^{8/3}}\beta_n^2\gamma_n^2 + \frac{32K^2\Sigma_w^2}{n^4}\Big) n \\
&\le \frac{\Sigma_w^2}{n} + \frac{24K^{1/3}\Sigma_w}{n^{4/3}}\beta_n\gamma_n + \frac{288K^{2/3}}{n^{5/3}}\beta_n^2\gamma_n^2 + \frac{16K\Sigma_w^2}{n^2}.
\end{aligned}$$

Finally, using (38), (39), and the previous inequality and recalling $\delta = n^{-9/2}$, Corollary 21, $\beta_n \leq C_\beta$, and $\gamma_n < 1.5$ from Appendix B.3 we have

$$\begin{aligned}
R_n(\mathcal{A}_{\text{MC-UCB}}) &= L_n(\mathcal{A}_{\text{MC-UCB}}) - \frac{\Sigma_w^2}{n} \\
&\leq \frac{24K^{1/3}\Sigma_w}{n^{4/3}}\beta_n\gamma_n + \frac{288K^{2/3}}{n^{5/3}}\beta_n^2\gamma_n^2 + \frac{16K\Sigma_w^2}{n^2} + \frac{KC_\xi}{n^{5/2}} \\
&\leq \frac{36K^{1/3}\Sigma_w C_\beta}{n^{4/3}} + \frac{648K^{2/3}C_\beta^2}{n^{5/3}} + \frac{\sqrt{K}C_\xi + 32K\Sigma_w^2}{2n^2} \\
&\leq \frac{36K^{1/3}\Sigma_w C_\beta}{n^{4/3}} + \frac{K^{2/3}(2058C_\beta^2 + 32\Sigma_w^2) + K^{1/6}C_\xi}{(2n)^{5/3}}.
\end{aligned}$$

This concludes the proof. ■

Remark 22 *Observe that in the proof of Proposition 8 and 9, we already bounded a linear combination of $\mathbb{E}[\mathbb{I}\{\xi\}/T_{k,n}]$ (leading to the desired rates), that is clearly upper bounded also by $\mathbb{E}[T_{k,n}]/\inf_\xi T_{k,n}^2$ appearing in both proofs above. Unfortunately, a reverse inequality does not directly hold, thus here we had to proceed in a more involved way leading to looser constants. If one could derive such a reverse inequality and then use the bounds on $\widetilde{R}_n(\mathcal{A}_{MC\text{-}UCB})$, that might give sharper constants also in the bounds on $R_n(\mathcal{A}_{MC\text{-}UCB})$.*

Appendix E. Bounds on the Cross Product-Terms

In this appendix, we prove Proposition 5 and 6 stating that the cross product-terms in (2) are 0 for symmetric distributions and decrease at polynomial rate in n in the general sub-Gaussian case.

E.1 Vanishing of the Terms for Symmetric Arm Distributions

Proof of Proposition 5

Step 1: Conditioning on a pair of numbers of pulls. Recall that $(\hat{s}_{k,t})_{k\leq K, t\leq n}$ are the unbiased empirical variances (see Equation 12). At each time step $t > 2K$, $\mathcal{A}_{\text{MC-UCB}}$ chooses k_t based on the values of $(B_{p,t})_{p\leq K}$, which depend on $\{T_{p,t-1}\}_{p\leq K}$ and $\{\hat{\sigma}_{p,t-1}\}_{p\leq K}$. Thus $\{T_{p,t}\}_{p\leq K}$ is a deterministic map of $\{T_{p,t-1}\}_{p\leq K}$ and $\{\hat{\sigma}_{p,t-1}\}_{p\leq K}$. Hence, by induction, each $T_{k,n}$ is a deterministic function of $\{\hat{\sigma}_{p,t}\}_{p\leq K, t<n}$, and so of $\{\hat{s}_{p,t}\}_{p\leq K, t\leq n}$, as well.

Now fix arms k,k' and $1 \leq s$,$s' \leq n$ such that $\mathbb{P}(T_{k,n} = s, T_{k',n} = s') > 0$. Then we have

$$\begin{aligned}
&\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)(\hat{\mu}_{k',n} - \mu_{k'})\big|T_{k,n} = s, T_{k',n} = s'\big] \\
&= \mathbb{E}\Big[\Big(\frac{1}{s}\sum_{t=1}^{s} X_{k,t} - \mu_k\Big)\Big(\frac{1}{s'}\sum_{t=1}^{s'} X_{k',t} - \mu_{k'}\Big)\Big|T_{k,n} = s, T_{k',n} = s'\Big] \qquad (42) \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\Big(\frac{1}{s}\sum_{t=1}^{s} X_{k,t} - \mu_k\Big)\Big(\frac{1}{s'}\sum_{t=1}^{s'} X_{k',t} - \mu_{k'}\Big)\Big|\{\hat{s}_{p,t}\}_{p\leq K, t\leq n}\Big]\Big|T_{k,n} = s, T_{k',n} = s'\Big].
\end{aligned}$$

Since the full sample sequences of the individual arms are independent, the sequences $(X_{k,1},\dots,X_{k,s})$ and $(X_{k',1},\dots,X_{k',s'})$ remain conditionally independent conditioning on $\{\hat{s}_{p,t}\}_{p\le K,t\le n}$. This leads to:

$$\begin{aligned}&\mathbb{E}\Big[\Big(\frac{1}{s}\sum_{t=1}^{s}X_{k,t}-\mu_k\Big)\Big(\frac{1}{s'}\sum_{t=1}^{s'}X_{k',t}-\mu_{k'}\Big)\Big|\{\hat{s}_{p,t}\}_{p\le K,t\le n}\Big] \qquad (43)\\ &\quad=\mathbb{E}\Big[\frac{1}{s}\sum_{t=1}^{s}X_{k,t}-\mu_k\Big|\{\hat{s}_{p,t}\}_{p\le K,t\le n}\Big]\mathbb{E}\Big[\frac{1}{s'}\sum_{t=1}^{s'}X_{k',t}-\mu_{k'}\Big|\{\hat{s}_{p,t}\}_{p\le K,t\le n}\Big].\end{aligned}$$

Step 2: For any $k\le K$ and $s\le n$, $\mathbb{E}\big[\frac{1}{s}\sum_{t=1}^{s}X_{k,t}-\mu_k\big|\{\hat{s}_{p,t}\}_{p\le K,t\le n}\big]=0$ a.s. We first state the following Lemma proven in Appendix F:

Lemma 23 *Let ν be a symmetric distribution on $\mathbb{R}$ around 0, $X=(X_1,\dots,X_n)$ be generated in an i.i.d. way according to ν, and $\hat{s}_2,\dots,\hat{s}_n$ are the unbiased empirical standard deviations given by (35). Then for $1\le t\le n$, $\mathbb{E}[X_t|\{\hat{s}_{t'}\}_{t'\le n}]=0$ a.s.*

As ν_k is symmetric, Lemma 23 applies to $X=(X_{k,1}-\mu_k,\dots,X_{k,n}-\mu_k)$ and $\{\hat{s}_{k,t}\}_{t\le n}$, that is,

$$\mathbb{E}\Big[\frac{1}{s}\sum_{t=1}^{s}X_{k,t}-\mu_k\Big|\{\hat{s}_{k,t}\}_{t\le n}\Big]=\frac{1}{s}\sum_{t=1}^{s}\mathbb{E}[X_{k,t}-\mu_k|\{\hat{s}_{k,t'}\}_{t'\le n}]=0 \qquad \text{a.s.}$$

By definition, $\{\hat{s}_{p,t}\}_{p\ne k,t\le n}$ is independent of $(X_{k,1},\dots,X_{k,s},\{\hat{s}_{k,t}\}_{t\le n})$, hence

$$\mathbb{E}\Big[\frac{1}{s}\sum_{t=1}^{s}X_{k,t}-\mu_k\Big|\{\hat{s}_{p,t}\}_{p\le K,t\le n}\Big]=\mathbb{E}\Big[\frac{1}{s}\sum_{t=1}^{s}X_{k,t}-\mu_k\Big|\{\hat{s}_{k,t}\}_{t\le n}\Big]=0 \quad \text{a.s.} \qquad (44)$$

Step 3: The cross product-terms $\mathbb{E}[(\hat{\mu}_{k,n}-\mu_k)(\hat{\mu}_{k',n}-\mu_{k'})]=0$. We combine (42), (43), and (44) to get in case of $\mathbb{P}(T_{k,n}=s,T_{k',n}=s')>0$

$$\mathbb{E}[(\hat{\mu}_{k,n}-\mu_k)(\hat{\mu}_{k',n}-\mu_{k'})|T_{k,n}=s,T_{k',n}=s']=\mathbb{E}[0\cdot 0|T_{k,n}=s,T_{k',n}=s']=0.$$

Conditioning on $\{T_{k,n}=s,T_{k',n}=s'\}$ and using the equation above

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

Taking the weighted sum over k and k' concludes the proof. ■

E.2 Bounds on the Terms for General Arm Distributions

The following lemma proven in Appendix F will be used for the proof:

Lemma 24 *Let X be a random variable. Let $(\Omega_u)_{u=1,\dots,p}$ be a partition of an event Ω' of the probability space. Let $a_u \in \mathbb{R}$ for $u = 1,\dots,p$, and $\underline{a} = \min_{1\le u\le p} a_u$, $\bar{a} = \max_{1\le u\le p} a_u$. We have*

$$\Big|\mathbb{E}\Big[X\sum_{u=1}^{p} a_u\mathbb{I}\{\Omega_u\}\Big]\Big| - \big|\underline{a}\mathbb{E}[X\mathbb{I}\{\Omega'\}]\big| \le \Big|\mathbb{E}\Big[X\sum_{u=1}^{p} a_u\mathbb{I}\{\Omega_u\}\Big] - \underline{a}\mathbb{E}[X\mathbb{I}\{\Omega'\}]\Big|$$
$$\le (\bar{a}-\underline{a})\mathbb{E}|X\mathbb{I}\{\Omega'\}\,|.$$

Proof of Proposition 6 For any given $k \neq q$, introduce

$$Z_{kq} \overset{\text{def}}{=} \left(\sum_{t=1}^{T_{k,n}}(X_{k,t}-\mu_k)\right)\left(\sum_{t=1}^{T_{q,n}}(X_{q,t}-\mu_q)\right) = T_{k,n}T_{q,n}(\hat{\mu}_{k,n}-\mu_k)(\hat{\mu}_{q,n}-\mu_q).$$

Then it suffices to bound $|w_k w_q \mathbb{E}[Z_{kq}/(T_{k,n}T_{q,n})]|$.

Step 1: $\mathbb{E}[Z_{kq}]=0$. Let $\mathcal{T}_{k,t} \overset{\text{def}}{=} \min\{s\ge 1 : T_{k,s}\ge t\}$, that is, that random time step when $\mathcal{A}_{\text{MC-UCB}}$ pulls arm k the t^{th} time. ($\mathcal{T}_{k,t}=\infty$ if k is not pulled t times.) Now

$$\begin{aligned}\mathbb{E}[Z_{kq}] &= \mathbb{E}\Big[\Big(\sum_{t=1}^{n}(X_{k,t}-\mu_k)\mathbb{I}\{T_{k,n}\ge t\}\Big)\Big(\sum_{t=1}^{n}(X_{q,t}-\mu_q)\mathbb{I}\{T_{q,n}\ge t\}\Big)\Big]\\ &= \sum_{t=1}^{n}\sum_{t'=1}^{n}\mathbb{E}\big[(X_{k,t}-\mu_k)(X_{q,t'}-\mu_q)\mathbb{I}\{T_{k,n}\ge t \wedge T_{q,n}\ge t'\}\big]\\ &= \sum_{t=1}^{n}\sum_{t'=1}^{n}\mathbb{E}\big[(X_{k,t}-\mu_k)(X_{q,t'}-\mu_q)\mathbb{I}\{T_{k,n}\ge t \wedge T_{q,n}\ge t' \wedge \mathcal{T}_{k,t}<\mathcal{T}_{q,t'}\}\big]\\ &\quad+\sum_{t=1}^{n}\sum_{t'=1}^{n}\mathbb{E}\big[(X_{k,t}-\mu_k)(X_{q,t'}-\mu_q)\mathbb{I}\{T_{k,n}\ge t \wedge T_{q,n}\ge t' \wedge \mathcal{T}_{k,t}>\mathcal{T}_{q,t'}\}\big].\end{aligned}$$

Fix any $1\le t,t'\le n$. Proposition 17 implies that $\{T_{k,n}\le t-1\}\in\mathcal{G}_{t-1}^{(k)}$ (defined in Proposition 17), and thus also $\{T_{k,n}\ge t\}\in\mathcal{G}_{t-1}^{(k)}$. $\mathcal{T}_{k,t}>\mathcal{T}_{q,t'}$ means that for some time step $s\ge t'$, $\{T_{q,s}\ge t'\}$, but $\{T_{k,s}<t\}$. Thus,

$$\{\mathcal{T}_{k,t}>\mathcal{T}_{q,t'}\} = \bigcup_{s=t'}^{\infty}\{T_{q,s}\ge t'\}\cap\{T_{k,s}<t\}.$$

Intersecting this by $\{T_{k,n}\ge t\}$ and noting that for $s\ge n$, $\{T_{k,s}<T_{k,n}\}=\emptyset$

$$\{\mathcal{T}_{k,t}>\mathcal{T}_{q,t'}\}\cap\{T_{k,n}\ge t\} = \bigcup_{s=t'}^{n-1}\{T_{q,s}\ge t'\}\cap\{T_{k,s}<t\le T_{k,n}\}.$$

Now, by Proposition 17 for any $s \leq n$, $\{T_{k,s} < t\} = \{T_{k,s} \leq t-1\} \in \mathcal{G}^{(k)}_{t-1}$. Moreover, on $\{T_{k,s} < t\}$, $T_{q,s}$ is $\mathcal{G}^{(k)}_{t-1}$-measurable, thus $\{T_{q,s} \geq t'\} \cap \{T_{k,s} < t\} \in \mathcal{G}^{(k)}_{t-1}$. Hence $\{\mathcal{T}_{k,t} > \mathcal{T}_{q,t'}\} \cap \{T_{k,n} \geq t\} \in \mathcal{G}^{(k)}_{t-1}$, as well. Observe that $T_{k,n} \geq t$ and $\mathcal{T}_{k,t} > \mathcal{T}_{q,t'}$ together imply $T_{q,n} \geq t'$, so $\mathbb{I}\big\{\mathcal{T}_{k,t} > \mathcal{T}_{q,t'} \wedge T_{k,n} \geq t \wedge T_{q,n} \geq t'\big\}$ is $\mathcal{G}^{(k)}_{t-1}$-measurable. Also $X_{q,t'}$ is obviously $\mathcal{G}^{(k)}_{t-1}$-measurable, while $X_{k,t}$ is independent of $\mathcal{G}^{(k)}_{t-1}$. Thus, conditioning on $\mathcal{G}^{(k)}_{t-1}$, we have

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

By summing for t,t' and repeating the same reasoning for the other term of $\mathbb{E}[Z_{kq}]$ with arm q, we obtain that $\mathbb{E}[Z_{kq}] = 0$.

Step 2: Bounding the terms on ξ^C. By Corollary 21 we have

$$\left|\mathbb{E}\left[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\mathbb{I}\{\xi^C\}\right]\right| \leq \frac{KC_\xi}{n^{5/2}} \qquad \text{and} \qquad \left|\mathbb{E}\left[Z_{kq}\mathbb{I}\{\xi^C\}\right]\right| \leq \frac{2KC_\xi}{n^{3/2}}, \tag{45}$$

implying, since $\mathbb{E}[Z_{kq}] = 0$ (Step 1), also

$$|\mathbb{E}[Z_{kq}\mathbb{I}\{\xi\}]| \leq \frac{2KC_\xi}{n^{3/2}}. \tag{46}$$

Step 3: Bounding the terms on ξ. We recall that under Assumption 1, $n \geq 4K$, and $\delta = n^{-9/2}$, combining Lemmata 1 (Equation 15) and 2 we have that $\mathcal{A}_{\text{MC-UCB}}$ run by β_n given by (11) satisfies on ξ for all arm p, $-\lambda_p M \leq T_{p,n} - T^*_{p,n} \leq M$, where

$$M \overset{\text{def}}{=} 4\min\left(\frac{3\beta_n}{\Sigma_w \lambda_{\min}^{3/2}}\sqrt{n} + K, K^{1/3}\frac{3\beta_n\gamma_n}{\Sigma_w}n^{2/3} + K\right)$$

and $\gamma_n = (\bar{\Sigma}/\beta_n + \sqrt{8})^{1/3}$ as in Lemma 2. Recalling $\beta_n \leq C_\beta$ and $\gamma_n < 1.5$ from Appendix B.3 M is upper bounded by $\min\big(B\sqrt{n}, An^{2/3}\big)$, where

$$B \overset{\text{def}}{=} \frac{12C_\beta}{\Sigma_w \lambda_{\min}^{3/2}} + 2\sqrt{K} \qquad \text{and} \qquad A \overset{\text{def}}{=} K^{1/3}\left(\frac{18C_\beta}{\Sigma_w} + 4^{1/3}\right).$$

Moreover, by (16) of Lemma 2,

$$T_{p,n} \geq \tfrac{(w_p n)^{2/3}}{\gamma_n^2} > 4(\underline{w}n)^{2/3}/9 = En^{2/3} \qquad \text{on } \xi,$$

where $\underline{w} \overset{\text{def}}{=} \min_k w_k$ and $E \overset{\text{def}}{=} 4\underline{w}^{2/3}/9 > 0$. Note that B displays a dependency on $\lambda_{\min}^{-1}$, but A and E do not. Summarizing these inequalities on $T_{p,n}$ we have

$$\begin{aligned}
T_{p,n} &\geq \max\Big(T^*_{p,n} - \lambda_p \min\big(B\sqrt{n}, An^{2/3}\big), En^{2/3}\Big) \overset{\text{def}}{=} \underline{T}_{p,n} \\
\text{and} \qquad T_{p,n} &\leq T^*_{p,n} + \min\big(B\sqrt{n}, An^{2/3}\big) \overset{\text{def}}{=} \bar{T}_{p,n}
\end{aligned}$$

on ξ. Note that using $n \geq 4K \geq 8$, $\Sigma_w^2 \leq c_1 \log(ec_2)$, $c_2 \geq 1$, and $\lambda_{\min} \leq 1/K$ it is easy to see that each $\bar{T}_{p,n} > 643$. Since now

$$\{\{T_{k,n} = t, T_{q,n} = t'\} \cap \xi : \underline{T}_{k,n} \leq t \leq \bar{T}_{k,n}, \underline{T}_{q,n} \leq t' \leq \bar{T}_{q,n}\}$$

is a partition of ξ, we have by Lemma 24

$$\begin{aligned}\Big|\mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\mathbb{I}\{\xi\}\Big]\Big| &= \Big|\mathbb{E}\Big[Z_{kq}\sum_{t=\underline{T}_{k,n}}^{\bar{T}_{k,n}}\sum_{t'=\underline{T}_{q,n}}^{\bar{T}_{q,n}}\frac{1}{tt'}\mathbb{I}\{\{T_{k,n} = t, T_{q,n} = t'\} \cap \xi\}\Big]\Big| \\ &\leq \mathbb{E}|Z_{kq}\mathbb{I}\{\xi\}\,|\Big(\frac{1}{\underline{T}_{k,n}\underline{T}_{q,n}} - \frac{1}{\bar{T}_{k,n}\bar{T}_{q,n}}\Big) + \frac{1}{\bar{T}_{k,n}\bar{T}_{q,n}}|\mathbb{E}[Z_{kq}\mathbb{I}\{\xi\}]|.\end{aligned}$$

Note now that by Cauchy-Schwarz's inequality

$$\mathbb{E}|Z_{kq}\mathbb{I}\{\xi\}\,| \leq \mathbb{E}|Z_{kq}| \leq \sqrt{\mathbb{E}\Big[\Big(\sum_{t=1}^{T_{k,n}}(X_{k,t} - \mu_k)\Big)^2\Big]\mathbb{E}\Big[\Big(\sum_{t=1}^{T_{q,n}}(X_{q,t} - \mu_q)\Big)^2\Big]}.$$

Using Corollary 19 the right-hand side is bounded by $\sqrt{\mathbb{E}T_{k,n}\sigma_k^2\mathbb{E}T_{q,n}\sigma_q^2}$. Since

$$\mathbb{E}T_{k,n} = \mathbb{E}[T_{k,n}\mathbb{I}\{\xi\}] + \mathbb{E}[T_{k,n}\mathbb{I}\{\xi^C\}] \leq \bar{T}_{k,n}\mathbb{P}(\xi) + 2Kn^2\delta \leq \bar{T}_{k,n} + 2Kn^{-5/2} \leq \bar{T}_{k,n} + \sqrt{2}/64$$

by definition of $\bar{T}_{k,n}$ and $\bar{T}_{k,n} > 643$, $\mathbb{E}T_{k,n} < (1 + \sqrt{2}/41152)\bar{T}_{k,n} < 1.01\bar{T}_{k,n}$. Similarly, $\mathbb{E}T_{q,n} < 1.01\bar{T}_{q,n}$. Thus we have $\mathbb{E}|Z_{kq}\mathbb{I}\{\xi\}\,| \leq 1.01\sigma_k\sigma_q\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}$. From this and (46), one gets

$$\begin{aligned}w_kw_q\Big|\mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\mathbb{I}\{\xi\}\Big]\Big| &\leq 1.01w_k\sigma_kw_q\sigma_q\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}\Big(\frac{1}{\underline{T}_{k,n}\underline{T}_{q,n}} - \frac{1}{\bar{T}_{k,n}\bar{T}_{q,n}}\Big) + \frac{2w_kw_q}{\bar{T}_{k,n}\bar{T}_{q,n}}\frac{KC_\xi}{n^{3/2}} \\ &\leq 1.01\frac{w_k\sigma_kw_q\sigma_q}{\underline{T}_{k,n}\underline{T}_{q,n}}\frac{\bar{T}_{k,n}\bar{T}_{q,n} - \underline{T}_{k,n}\underline{T}_{q,n}}{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}} + \frac{1.3KC_\xi}{10^6n^{3/2}}.\end{aligned}$$

Now for n large enough (compared to K, c_1, $\log c_2$, $1/\Sigma_w$, and $\log n$), $n \geq 8A^3$ (i.e., $An^{2/3} \leq n/2$) holds. Thus

$$\underline{T}_{p,n} \geq T^*_{p,n} - A\lambda_pn^{2/3} = \lambda_p(n - An^{2/3})$$

implies also $\frac{w_p\sigma_p}{\underline{T}_{p,n}} \leq \frac{\Sigma_w}{n - An^{2/3}} \leq 2\frac{\Sigma_w}{n}$ for any arm p. This leads to the bound

$$w_kw_q\Big|\mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\mathbb{I}\{\xi\}\Big]\Big| \leq 4.04\frac{\Sigma_w^2}{n^2}\frac{\bar{T}_{k,n}\bar{T}_{q,n} - \underline{T}_{k,n}\underline{T}_{q,n}}{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}} + \frac{1.3KC_\xi}{10^6n^{3/2}}. \tag{47}$$

Step 4: problem-dependent upper bound. We deduce that

$$\begin{aligned}\frac{\bar{T}_{k,n}\bar{T}_{q,n}-\underline{T}_{k,n}\underline{T}_{q,n}}{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}} &\leq \frac{\big(n\lambda_k+B\sqrt{n}\big)\big(n\lambda_q+B\sqrt{n}\big)-\big(n\lambda_k-B\lambda_k\sqrt{n}\big)\big(n\lambda_q-B\lambda_q\sqrt{n}\big)}{\sqrt{n\lambda_k n\lambda_q}}\\
&=\frac{B(\lambda_k+\lambda_q+2\lambda_k\lambda_q)n\sqrt{n}+B^2(1-\lambda_k\lambda_q)n}{n\sqrt{\lambda_k\lambda_q}}\\
&\leq B\sqrt{n}\left(\frac{1+B/\sqrt{n}}{\sqrt{\lambda_k\lambda_q}}+2\sqrt{\lambda_k\lambda_q}\right)\\
&\leq B\left(\frac{1+B/\sqrt{8}}{\lambda_{\min}}+1\right)\sqrt{n}\end{aligned}$$

using $n\geq 4K\geq 8$ and $\lambda_k\lambda_q\leq 1/4$. Thus, we have from this and (47)

$$w_kw_q\Big|\mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\mathbb{I}\{\xi\}\Big]\Big|\leq 5B\left(\frac{1+B/\sqrt{8}}{\lambda_{\min}}+1\right)\frac{\Sigma_w^2}{n^{3/2}}+\frac{1.3KC_\xi}{10^6n^{3/2}}=\frac{C_1+1.3KC_\xi/10^6}{n^{3/2}},$$

where $C_1\overset{\mathrm{def}}{=}5B((1+B/\sqrt{8})/\lambda_{\min}+1)\Sigma_w^2$.

Finally, using (45), we have

$$\begin{aligned}\Big|w_kw_q\mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\Big]\Big| &\leq w_kw_q\Big|\mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\mathbb{I}\{\xi\}\Big]\Big|+w_kw_q\Big|\mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n}T_{q,n}}\mathbb{I}\{\xi^C\}\Big]\Big|\\
&\leq\frac{C_1+1.3KC_\xi/10^6}{n^{3/2}}+\frac{KC_\xi}{4n^{5/2}}\\
&\leq\frac{C_1+(1.3K/10^6+1/16)C_\xi}{n^{3/2}},\end{aligned}$$

where C_1 and C_ξ depend only polynomially on $\log n$, $\lambda_{\min}^{-1}$, K, Σ_w, c_1, and $\log c_2$. This concludes the proof for the problem-dependent bound.

Step 4': problem-independent upper bound. Using $\bar{T}_{k,n}\geq\underline{T}_{k,n}\geq En^{2/3}$, which implies that $\bar{T}_{k,n}\geq\max(\lambda_kn,En^{2/3})$, we deduce that

$$\begin{aligned}\frac{\bar{T}_{k,n}\bar{T}_{q,n}-\underline{T}_{k,n}\underline{T}_{q,n}}{\sqrt{\bar{T}_{k,n}\bar{T}_{q,n}}} &\leq\frac{\big(n\lambda_k+An^{2/3}\big)\big(n\lambda_q+An^{2/3}\big)-\big(n\lambda_k-A\lambda_kn^{2/3}\big)\big(n\lambda_q-A\lambda_qn^{2/3}\big)}{\sqrt{\max\big(\lambda_kn,En^{2/3}\big)\max\big(\lambda_qn,En^{2/3}\big)}}\\
&=\frac{A(\lambda_k+\lambda_q+2\lambda_k\lambda_q)nn^{2/3}+A^2(1-\lambda_k\lambda_q)n^{4/3}}{\sqrt{\max\big(\lambda_k\lambda_qn^2,E\max(\lambda_k,\lambda_q)nn^{2/3},E^2n^{4/3}\big)}}\\
&\leq A\left[\frac{(\lambda_k+\lambda_q)n^{5/3}}{\sqrt{E\max(\lambda_k,\lambda_q)n^{5/3}}}+\frac{2\lambda_k\lambda_qn^{5/3}}{\sqrt{\lambda_k\lambda_q}n}+\frac{An^{4/3}}{En^{2/3}}\right]\\
&\leq An^{5/6}\left[\frac{\sqrt{\lambda_k+\lambda_q}}{\sqrt{E/2}}+\frac{2\sqrt{\lambda_k\lambda_q}}{n^{1/6}}+\frac{A}{En^{1/6}}\right]\\
&\leq\frac{A}{\sqrt{2}}\Big(\frac{2}{\sqrt{E}}+1+\frac{A}{E}\Big)n^{5/6}\end{aligned}$$

using $n \geq 4K \geq 8$ and $\lambda_k \lambda_q \leq 1/4$. Thus, we have from this and (47) that

$$w_k w_q \Big| \mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n} T_{q,n}} \mathbb{I}\{\xi\} \Big] \Big| \leq 2.02\sqrt{2}A\Big(\frac{2}{\sqrt{E}} + 1 + \frac{A}{E} \Big) \frac{\Sigma_w^2}{n^{7/6}} + \frac{1.3 K C_\xi}{10^6 n^{3/2}} \leq \frac{C_2 + 9K^{2/3} C_\xi / 10^7}{n^{7/6}},$$

where $C_2 = 3A(2/\sqrt{E} + 1 + A/E)\Sigma_w^2$.

Finally, using (45), we have

$$\begin{aligned} \Big| w_k w_q \mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n} T_{q,n}} \Big] \Big| &\leq w_k w_q \Big| \mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n} T_{q,n}} \mathbb{I}\{\xi\} \Big] \Big| + w_k w_q \Big| \mathbb{E}\Big[\frac{Z_{kq}}{T_{k,n} T_{q,n}} \mathbb{I}\{\xi^C\} \Big] \Big| \\ &\leq \frac{C_2 + 9K^{2/3} C_\xi / 10^7}{n^{7/6}} + \frac{K C_\xi}{4 n^{5/2}} \\ &\leq \frac{C_2 + (9K^{2/3}/10^7 + 1/32) C_\xi)}{n^{7/6}}, \end{aligned}$$

where C_2 and C_ξ depend only polynomially on $\log n$, K, Σ_w, c_1, $\log c_2$, and $1/\underline{w}$. This concludes the proof for the problem-independent bound. ■

Appendix F. Proofs of Technical Lemmata

Proof of Lemma 11 Using that $\log(c_2/\delta) \geq 0$

$$\begin{aligned} \mathbb{E}\left[|X - \mu|^2 \mathbb{I}\{A\} \right] &= \int_0^\infty \mathbb{P}\big(|X-\mu|^2 > \epsilon, A \big)\, d\epsilon \\ &\leq \int_0^{c_1 \log(c_2/\delta)} \mathbb{P}(A)\, d\epsilon + \int_{c_1 \log(c_2/\delta)}^\infty \mathbb{P}\big(|X-\mu|^2 > \epsilon \big)\, d\epsilon \\ &\leq \delta c_1 \log(c_2/\delta) + \int_{c_1 \log(c_2/\delta)}^\infty c_2 e^{-\epsilon/c_1}\, d\epsilon = \delta c_1 \log(e c_2/\delta). \end{aligned}$$

■

Proof of Lemma 15 Using (9) for $\epsilon^2 = c_1 \log(c_2/\delta) (> 0)$ we have

$$\mathbb{P}(A^C) \leq c_2 e^{-c_1 \log(c_2/\delta)/c_1} = \delta \qquad \text{and} \qquad \mathbb{P}(A) \geq 1 - \delta > 0, \tag{48}$$

so $\mathrm{Var}[X|A]$ and also $\tilde{\mu} \overset{\text{def}}{=} \mathbb{E}[X|A] = \mathbb{E}[X \mathbb{I}\{A\}]/\mathbb{P}(A)$ make sense. If $\mathbb{P}(A) = 1$ then $\tilde{\sigma} = \sigma$, and the claim follows. Now assume $\mathbb{P}(A) < 1$. Since $\mathbb{E}[|X-\mu|^2 | A^C] \geq c_1 \log(c_2/\delta) \geq \mathbb{E}[|X-\mu|^2|A]$, we have

$$\sigma^2 = \mathbb{E}[|X-\mu|^2] = \mathbb{E}[|X-\mu|^2|A^C]\mathbb{P}(A^C) + \mathbb{E}[|X-\mu|^2|A]\mathbb{P}(A) \geq \mathbb{E}[|X-\mu|^2|A]. \tag{49}$$

Moreover,

$$\tilde{\sigma}^2 = \mathbb{E}[|X - \tilde{\mu}|^2 | A] = \mathbb{E}[|X-\mu|^2|A] - |\mu - \tilde{\mu}|^2,$$

and thus

$$\sigma^2 - \tilde{\sigma}^2 = \sigma^2 - \mathbb{E}[|X-\mu|^2|A] + |\mu - \tilde{\mu}|^2 \geq 0 \tag{50}$$

by (49). But (49) implies also that

$$\begin{aligned}\sigma^2 - \mathbb{E}[|X-\mu|^2|A] &= \frac{\sigma^2\mathbb{P}(A) - \mathbb{E}[|X-\mu|^2\mathbb{I}\{A\}]}{\mathbb{P}(A)} = \frac{\mathbb{E}[|X-\mu|^2\mathbb{I}\{A^C\}] - \sigma^2\mathbb{P}(A^C)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{E}[(|X-\mu|^2 - \sigma^2)\mathbb{I}\{A^C\}]}{\mathbb{P}(A)}. \end{aligned}\tag{51}$$

Using that $\delta \leq 1/e$ and Lemma 11 imply $c_1 \log(c_2/\delta) \geq c_1 \log(ec_2) \geq \sigma^2$, we have

$$\begin{aligned}&\mathbb{E}\big[(|X-\mu|^2 - \sigma^2)\mathbb{I}\{A^C\}\big] = \int_0^\infty \mathbb{P}(|X-\mu|^2 - \sigma^2 > \epsilon', A^C)\, d\epsilon' = \int_{\sigma^2}^\infty \mathbb{P}(|X-\mu|^2 > \epsilon, A^C)\, d\epsilon \\ &= \int_{\sigma^2}^{c_1\log(c_2/\delta)} \mathbb{P}(A^C)\, d\epsilon + \int_{c_1\log(c_2/\delta)}^\infty \mathbb{P}(|X-\mu|^2 > \epsilon)\, d\epsilon \\ &\leq \delta(c_1\log(c_2/\delta) - \sigma^2) + \int_{c_1\log(c_2/\delta)}^\infty c_2 e^{-\epsilon/c_1}\, d\epsilon \qquad \text{(by Equations 48 and 9)} \\ &= \delta c_1 \log(c_2/\delta) - \delta\sigma^2 + c_1c_2e^{-c_1\log(c_2/\delta)/c_1} = \delta c_1 \log(ec_2/\delta) - \delta\sigma^2. \end{aligned}$$

This, (51), and (48) imply

$$\sigma^2 - \mathbb{E}[|X-\mu|^2|A] \leq \delta\frac{c_1\log(ec_2/\delta) - \sigma^2}{1-\delta}. \tag{52}$$

For the last term of (50), noticing that $\mathbb{E}[X\mathbb{I}\{A^C\}] + \mathbb{E}[X\mathbb{I}\{A\}] = \mu$ we have

$$\begin{aligned}|\mu - \tilde{\mu}| &= \left|\frac{\mu\mathbb{P}(A) - \mathbb{E}[X\mathbb{I}\{A\}]}{\mathbb{P}(A)}\right| = \frac{\big|\mathbb{E}[X\mathbb{I}\{A^C\}] - \mu\mathbb{P}(A^C)\big|}{\mathbb{P}(A)} = \frac{\big|\mathbb{E}[(X-\mu)\mathbb{I}\{A^C\}]\big|}{\mathbb{P}(A)} \\ &\leq \frac{\sqrt{\mathbb{E}[|X-\mu|^2]\mathbb{E}[\mathbb{I}\{A^C\}]}}{\mathbb{P}(A)} \qquad \text{(by Cauchy-Schwarz's inequality)} \\ &= \frac{\sigma\sqrt{\mathbb{P}(A^C)}}{\mathbb{P}(A)} \leq \frac{\sigma\sqrt{\delta}}{1-\delta} \end{aligned}\tag{53}$$

using again (48). From (50), (52), and (53), we derive

$$\sigma^2 - \tilde{\sigma}^2 \leq \delta\frac{c_1\log(ec_2/\delta) - \sigma^2}{1-\delta} + \frac{\delta\sigma^2}{(1-\delta)^2} \leq c_1\delta\frac{\log(ec_2/\delta)}{(1-\delta)^2}.$$

Since $(\sigma - \tilde{\sigma})^2 \leq (\sigma + \tilde{\sigma})(\sigma - \tilde{\sigma}) = \sigma^2 - \tilde{\sigma}^2$, the claim follows. ■

Proof of Lemma 23 Denote $(\hat{s}_2, \dots, \hat{s}_n)$ by $\hat{S}(X)$. Then $\hat{S}(X) = \hat{S}(-X)$ holds due to the quadratic form of the empirical variances. Thus, by the symmetry of ν,

$$\mathbb{E}[X_t|\hat{S}(X)] = \mathbb{E}[-X_t|\hat{S}(-X)] = -\mathbb{E}[X_t|\hat{S}(X)] \quad \text{a.s.},$$

implying $\mathbb{E}[X_t|\{\hat{s}_{t'}\}_{t' \leq n}] = \mathbb{E}[X_t|\hat{S}(X)] = 0$ a.s. ■

Proof of Lemma 24 By the definition of $\bar{a}$ and $\underline{a}$,

$$X \sum_{u=1}^{p} a_u \mathbb{I}\{\Omega_u\} \leq X\mathbb{I}\{X \geq 0\}\, \bar{a}\mathbb{I}\big\{\Omega'\big\} + X\mathbb{I}\{X < 0\}\, \underline{a}\mathbb{I}\big\{\Omega'\big\}\,.$$

This implies

$$\begin{aligned}
\mathbb{E}\Big[X \sum_{u=1}^{p} a_u \mathbb{I}\{\Omega_u\}\Big] &\leq \mathbb{E}\Big[X\mathbb{I}\{X \geq 0\}\, \bar{a}\mathbb{I}\big\{\Omega'\big\} + X\mathbb{I}\{X < 0\}\, \underline{a}\mathbb{I}\big\{\Omega'\big\}\Big] \\
&= \mathbb{E}\Big[(\bar{a} - \underline{a})X\mathbb{I}\{X \geq 0\}\, \mathbb{I}\big\{\Omega'\big\} + \underline{a}X(\mathbb{I}\{X < 0\} + \mathbb{I}\{X \geq 0\})\mathbb{I}\big\{\Omega'\big\}\Big] \\
&= (\bar{a} - \underline{a})\mathbb{E}\Big[X\mathbb{I}\{X \geq 0\}\, \mathbb{I}\big\{\Omega'\big\}\Big] + \underline{a}\mathbb{E}[X\mathbb{I}\big\{\Omega'\big\}] \\
&\leq (\bar{a} - \underline{a})\mathbb{E}|X\mathbb{I}\big\{\Omega'\big\}| + \underline{a}\mathbb{E}[X\mathbb{I}\big\{\Omega'\big\}].
\end{aligned}$$

By applying the inequality above for $-X$ we have

$$\mathbb{E}\Big[X \sum_{u=1}^{p} a_u \mathbb{I}\{\Omega_u\}\Big] \geq -(\bar{a} - \underline{a})\mathbb{E}|X\mathbb{I}\big\{\Omega'\big\}| + \underline{a}\mathbb{E}[X\mathbb{I}\big\{\Omega'\big\}].$$

Those two inequalities lead to the second inequality of the lemma, while the first one follows from the triangle inequality. ■

References

A. Antos. *Performance Limits of Nonparametric Estimators.* PhD thesis, Technical University of Budapest, May 1999. URL `http://www.cs.bme.hu/~antos/ps/dr.pdf`.

A. Antos, V. Grover, and Cs. Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29–30):2712–2728, June 17 2010. doi: 10.1016/j.tcs.2010.04.007. Special Issue for ALT 2008. Available online 10 April 2010.

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

K.B. Athreya and S.N. Lahiri. *Measure Theory and Probability Theory.* Springer, 2006.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In S. Dasgupta and A. Klivans, editors, *Proceedings of the* 22^{nd} *Annual Conference on Learning Theory*, pages 217–226. Omnipress, June 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.

J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proceedings of the* 23rd *Annual Conference on Learning Theory*, pages 41–53, 2010.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002. ISSN 0885-6125.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, December 12, 2012. doi: 10.1561/2200000024.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, April 22, 2011. ISSN 03043975. doi: 10.1016/j.tcs.2010.12.059.

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

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

A. Carpentier and R. Munos. Minimax number of strata for online stratified sampling given noisy samples. In N.H. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann, editors, *Proceedings of the* 23rd *International Conference, Algorithmic Learning Theory 2012*, volume 7568 of *LNCS/LNAI*, pages 229–244, Berlin, Heidelberg, 2012. Springer-Verlag.

A. Carpentier, A. Lazaric, M. Ghavamzadeh, R. Munos, and P. Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In J. Kivinen, C. Szepesvári, E. Ukkonen, and Th. Zeugmann, editors, *Proceedings of the* 22^{nd} *International Conference, Algorithmic Learning Theory 2011*, volume 6925 of *LNCS/LNAI*, pages 189–203, Berlin, Heidelberg, 2011. Springer-Verlag.

A. Carpentier, A. Lazaric, M. Ghavamzadeh, R. Munos, P. Auer, and A. Antos. Upper-confidence-bound algorithms for active learning in multi-armed bandits. *ArXiv e-prints*, July 2015. URL `http://arxiv.org/abs/1507.04523`. Technical Report.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge Univ Press, 2006. ISBN 0521841089.

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

P. Etoré, G. Fort, B. Jourdain, and É. Moulines. On adaptive stratification. *Annals of Operations Research*, 189(1):127–154, September 2011. doi: 10.1007/s10479-009-0638-9. Published online: November 21, 2009.

P. Glasserman. *Monte Carlo Methods in Financial Engineering.* Springer Verlag, 2004. ISBN 0387004513.

V. Grover. Active learning and its application to heteroscedastic problems. Master's thesis, Department of Computing Science, Univ. of Alberta, Edmonton, AB, Canada, 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.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the* 22^{nd} *Annual Conference on Learning Theory*, pages 115–124, 2009.

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