Upper-Confidence-Bound Algorithms for Active Learning in Multi-Armed Bandits

Alexandra Carpentier*, Alessandro Lazaric*, Mohammad Ghavamzadeh*, Rémi Munos*, Peter Auer **, András Antos ***

(*) SequeL team, INRIA Lille - Nord Europe, Team SequeL, France (**) University of Leoben, Franz-Josef-Strasse 18, 8700 Leoben, Austria (***) Budapest University of Technology and Economics, Műegyetem rkp. 3, 1111 Budapest, Hungary

Abstract

In this paper, we study the problem of estimating uniformly well the mean values of several distributions given a finite budget of samples. If the variance of the distributions were known, one could design an optimal sampling strategy by collecting a number of independent samples per distribution that is proportional to their variance. However, in the more realistic case where the distributions are not known in advance, one needs to design adaptive sampling strategies in order to select which distribution to sample from according to the previously observed samples. We describe two strategies based on pulling the distributions a number of times that is proportional to a high-probability upper-confidence-bound on their variance (built from previous observed samples) and report a finite-sample performance analysis on the excess estimation error compared to the optimal allocation. We show that the performance of these allocation strategies depends not only on the variances but also on the full shape of the distributions.

Keywords: Bandit Theory, Active Learning

1. Introduction

Consider a marketing problem where the objective is to estimate the potential impact of several new products or services. A common approach to this problem is to design active online polling systems, where at each time a product is presented (e.g., via a web banner on Internet) to random customers from a population of interest, and feedbacks are collected (e.g., whether the customer clicks on the ad or not) and used to estimate the average preference of all the products. It is often the case that some products have a general consensus of opinion (low variance) while others have a large variability (high variance). While in the former case very few votes would be enough to have an accurate estimate of the value of the product, in the latter the system should present the product to more customers in order to achieve the same accuracy. Since the variability of the opinions for different products is not known in advance, the objective is to design an active strategy that selects which product to display at each time step in order to estimate the values of all the products uniformly well.

The problem of online polling can be seen as an online allocation problem with several options, where the accuracy of the estimation of the quality of each option depends on the quantity of the resources allocated to it and also on some (initially unknown) intrinsic variability of the option. This general problem is closely related to the problems of active learning [8, 6], sampling and Monte-Carlo methods [10], and optimal experimental design [11, 7]. A particular instance of this problem is introduced in [1] as an active learning problem in the framework of stochastic multi-armed bandits. More precisely, the problem is modeled as a repeated game between a learner and a stochastic environment, defined by a set of K unknown distributions $\{\nu_k\}_{k=1}^K$, where at each round t, the learner selects an action (or arm) k_t and as a consequence receives a random sample from ν_{k_t} (independent of the past samples). Given a total budget of n samples, the goal is to define an allocation strategy over arms so as to estimate their expected values uniformly well. Note that if

the variances $\{\sigma_k^2\}_{k=1}^K$ of the arms were initially known, the optimal allocation strategy would be to sample the arms proportionally to their variances, or more accurately, proportionally to $\lambda_k = \sigma_k^2 / \sum_j \sigma_j^2$. However, since the distributions are initially unknown, the learner should follow an active allocation strategy which adapts its behavior as samples are collected. The performance of this strategy is measured by its regret (defined precisely by Equation 4) that is the difference between the maximal expected quadratic estimation error of the algorithm and the maximal expected error of the optimal allocation.

Antos et al. [1] presented an algorithm, called GAFS-MAX, that allocates samples proportionally to the empirical variances of the arms, while imposing that each arm should be pulled at least \sqrt{n} times (to guarantee good estimation of the true variances), where n is the total budget of pulls. They proved that for large enough n, the regret of their algorithm scales with $\tilde{O}(n^{-3/2})$ and conjectured that this rate is optimal. However, the performance displays both an implicit (in the condition for large enough n) and explicit (in the regret bound) dependency on the inverse of the smallest optimal allocation proportion, i.e., $\lambda_{\min} = \min_k \lambda_k$. This suggests that the algorithm is expected to have a poor performance whenever an arm has a very small variance compared to the others. Whether this dependency is due to the analysis of GAFS-MAX, to the specific class of algorithms, or to an intrinsic characteristic of the problem is an interesting open question. One of the main objectives of this paper is to investigate this issue and identify under which conditions this dependency can be avoided. Our main contributions and findings are as follows:

- We introduce two new algorithms based on upper-confidence-bounds (UCB) on the variance.
- The first algorithm, called CH-AS, is based on Chernoff-Hoeffding's bound, whose regret has the rate $\tilde{O}(n^{-3/2})$ and inverse dependency on λ_{\min} , similar to GAFS-MAX. The main differences are: the bound for CH-AS holds for any n (and not only for large enough n), multiplicative constants are made explicit, and finally, the proof is simpler and relies on very simple tools.
- The second algorithm, called B-AS, uses a sharper inequality than CH-AS, and has a better performance (in terms of the number of pulls) in targeting the optimal allocation strategy without any dependency on λ_{\min} . However, moving from the number of pulls to the regret causes the inverse dependency on λ_{\min} to appear in the bound again. We show that this might be due to specific shape of the distributions $\{\nu_k\}_{k=1}^K$ and derive a regret bound independent of λ_{\min} for the case of Gaussian arms.
- We show empirically that while the performance of CH-AS depends on λ_{min} in the case of Gaussian arms, this dependence does not exist for B-AS and GAFS-MAX, as they perform well in this case. This suggests that 1) it is not possible to remove λ_{min} from the regret bound of CH-AS, independent of the arms' distributions, and 2) GAFS-MAX's analysis could be improved along the same line as the proof of B-AS for the Gaussian arms. We also report experiments providing insights on the (somehow unexpected) fact that the full shapes of the distributions, and not only their variances, impact the regret of these algorithms.

2. Preliminaries

The allocation problem studied in this paper is formalized as the standard K-armed stochastic bandit setting, where each arm k = 1, ..., K is characterized by a distribution ν_k with mean μ_k and non-zero variance $\sigma_k^2 > 0$. At each round $t \ge 1$, the learner (algorithm \mathcal{A}) selects an arm k_t and receives a sample drawn from ν_{k_t} independently of the past. The objective is to estimate the mean values of all the arms uniformly well given a total budget of n pulls. An adaptive algorithm defines its allocation strategy as a function of the samples observed in the past (i.e., at time t, the selected arm k_t is a function of all the observations up to time t-1). After n rounds and observing $T_{k,n} = \sum_{t=1}^{n} \mathbb{I}\{k = k_t\}$ samples from each

¹The notation $u_n = \tilde{O}(v_n)$ means that there exist C > 0 and $\alpha > 0$ such that $u_n \leq C(\log n)^{\alpha}v_n$ for sufficiently large n.

arm k, the algorithm \mathcal{A} returns the empirical estimates $\hat{\mu}_{k,n} = \frac{1}{T_{k,n}} \sum_{t=1}^{T_{k,n}} X_{k,t}$, where $X_{k,t}$ denotes the sample received when we pull arm k for the t-th time. The accuracy of the estimation of each arm k is measured

received when we pull arm k for the t-th time. The accuracy of the estimation of each arm k is measured according to its expected squared estimation error, or loss

$$L_{k,n} = \mathbb{E}_{(\nu_i)_{i \le K}} \left[(\mu_k - \hat{\mu}_{k,n})^2 \right]. \tag{1}$$

The global performance or loss of \mathcal{A} is defined as the worst loss of the arms

$$L_n(\mathcal{A}) = \max_{1 \le k \le K} L_{k,n} . \tag{2}$$

If the variance of the arms were known in advance, one could design an optimal static allocation (i.e., the number of pulls does not depend on the observed samples) by pulling the arms proportionally to their variances. In the case of static allocation, if an arm k is pulled a fixed number of times $T_{k,n}^*$, its loss is computed as²

$$L_{k,n} = \frac{\sigma_k^2}{T_{k,n}^*} \,. \tag{3}$$

By choosing $T_{k,n}^*$ so as to minimize L_n under the constraint that $\sum_{k=1}^K T_{k,n}^* = n$, the optimal static allocation strategy \mathcal{A}^* pulls each arm k (up to rounding effects) $T_{k,n}^* = \frac{\sigma_k^2 n}{\sum_{k=1}^K \sigma_i^2}$ times, and achieves a global performance $L_n(\mathcal{A}^*) = \Sigma/n$, where $\Sigma = \sum_{i=1}^K \sigma_i^2$. We denote by $\lambda_k = \frac{T_{k,n}^*}{n} = \frac{\sigma_k^2}{\Sigma}$, the optimal allocation proportion for arm k, and by $\lambda_{\min} = \min_{1 \le k \le K} \lambda_k$, the smallest such proportion.

In our setting where the variances of the arms are not known in advance, the exploration-exploitation trade-off is inevitable: an adaptive algorithm \mathcal{A} should estimate the variances of the arms (exploration) at the same time as it tries to sample the arms proportionally to these estimates (exploitation). In order to measure how well the adaptive algorithm \mathcal{A} performs, we compare its performance to that of the optimal allocation algorithm \mathcal{A}^* , which requires the knowledge of the variances of the arms. For this purpose, we define the notion of regret of an adaptive algorithm \mathcal{A} as the difference between its loss $L_n(\mathcal{A})$ and the optimal loss $L_n(\mathcal{A}^*)$, i.e.,

$$R_n(\mathcal{A}) = L_n(\mathcal{A}) - L_n(\mathcal{A}^*). \tag{4}$$

It is important to note that unlike the standard multi-armed bandit problems, we do not consider the notion of cumulative regret, and instead, use the excess-loss suffered by the algorithm at the end of the n rounds. This notion of regret is closely related to the *pure exploration* setting (e.g., [3, 5]). An interesting feature that is shared between this setting and the problem of active learning considered in this paper is that good strategies should play all the arms as a linear function of n. This is in contrast with the standard stochastic bandit setting, at which the sub-optimal arms should be played logarithmically in n.

In [1], the authors provide an algorithm called GAFS-MAX and they prove that its regret is such that $R_n(\mathcal{A}_{GAFS-MAX}) = \tilde{O}(n^{-3/2})$ for a large enough budget n that depends on λ_{\min} . Also, the \tilde{O} depends on λ_{\min} . The smaller λ_{\min} , the larger n needs to be so that the bound in $\tilde{O}(n^{-3/2})$ holds, and also the larger the constant in the \tilde{O} .

3. Allocation Strategy Based on Chernoff-Hoeffding UCB

The first algorithm, called *Chernoff-Hoeffding Allocation Strategy* (CH-AS), is based on a Chernoff-Hoeffding high-probability bound on the difference between the estimated and true variances of the arms. Each arm is simply pulled proportionally to an upper-confidence-bound (UCB) on its variance. This algorithm deals with the exploration-exploitation trade-off by pulling more the arms with higher estimated variances or higher uncertainty in these estimates.

 $^{^{2}}$ This equality does not hold when the number of pulls is random, e.g., in adaptive algorithms where the strategy depends on the random observed samples.

```
Input: parameter \delta
Initialize: Pull each arm twice
for t = 2K + 1, \ldots, n do

Compute B_{k,t} = \frac{1}{T_{k,t-1}} \left( \hat{\sigma}_{k,t-1}^2 + 3\sqrt{\frac{\log(1/\delta)}{2T_{k,t-1}}} \right) for each arm 1 \le k \le K
Pull an arm k_t \in \arg\max_{1 \le k \le K} B_{k,t}
end for
Output: \hat{\mu}_{k,n} for all arms 1 \le k \le K
```

Figure 1: The pseudo-code of the CH-AS algorithm, with $\hat{\sigma}_{k,t}^2$ computed as in Equation 5.

3.1. The CH-AS Algorithm

The CH-AS algorithm \mathcal{A}_{CH} in Fig. 1 takes a confidence parameter δ as input and after n pulls returns an empirical mean $\hat{\mu}_{k,n}$ for each arm k. At each time step t, i.e., after having pulled arm k_t , the algorithm computes the empirical mean $\hat{\mu}_{k,t}$ and variance $\hat{\sigma}_{k,t}^2$ of each arm k as³

$$\hat{\mu}_{k,t} = \frac{1}{T_{k,t}} \sum_{i=1}^{T_{k,t}} X_{k,i} \quad \text{and} \quad \hat{\sigma}_{k,t}^2 = \frac{1}{T_{k,t}} \sum_{i=1}^{T_{k,t}} X_{k,i}^2 - \hat{\mu}_{k,t}^2 ,$$
 (5)

where $X_{k,i}$ is the *i*-th sample of ν_k and $T_{k,t}$ is the number of pulls⁴ allocated to arm k up to time t. After pulling each arm twice (rounds t = 1 to 2K), from round t = 2K + 1 on, the algorithm computes the $B_{k,t}$ values based on a Chernoff-Hoeffding's bound on the variances of the arms:

$$B_{k,t} = \frac{1}{T_{k,t-1}} \left(\hat{\sigma}_{k,t-1}^2 + 3\sqrt{\frac{\log(1/\delta)}{2T_{k,t-1}}} \right),$$

and then pulls the arm k_t with the largest $B_{k,t}$. This bound relies on the assumption that the distributions $\{\nu_k\}_{k=1}^K$ are supported [0,1].

Note that actually $\hat{\mu}_{k,t}$, $\hat{\sigma}_{k,t}$, $B_{k,t}$, k_t , and $T_{k,t}$ depend on the arm index (except for k_t), on the time step $t \leq n$, but also, either in a direct or in an indirect way (through the mechanism of the algorithm) on the budget n and on δ which will be chosen as a function of the budget n. However, since we consider most of the time a fixed budget n and thus a fixed δ , we conserve this notation in order to have lighter notations.

3.2. Regret Bound and Discussion

Before reporting a regret bound for the CH-AS algorithm, we first analyze its performance in targeting the optimal allocation strategy in terms of the number of pulls. As it will be discussed later, the distinction between the performance in terms of the number of pulls and the regret will allow us to stress the potential dependency of the regret on the distribution of the arms (see Section 4.3).

Lemma 1. Assume that the distributions $\{\nu_k\}_{k=1}^K$ are supported on [0,1] and let $\delta > 0$. Define the event

$$\xi_{K,n}^{CH}(\delta) = \bigcap_{\substack{1 \le k \le K \\ 1 \le t \le n}} \left\{ \left| \left(\frac{1}{t} \sum_{i=1}^{t} X_{k,i}^2 - \left(\frac{1}{t} \sum_{i=1}^{t} X_{k,i} \right)^2 \right) - \sigma_k^2 \right| \le 3\sqrt{\frac{\log(1/\delta)}{2t}} \right\}.$$

³Notice that this is a biased estimator of the variance even if the numbers of pulls $T_{k,t}$ were not random.

⁴An accurate notation for this should be $T_{k,t,n}$ since the number of pulls at time t depends also on n. However, for the sake of concision, we note $T_{k,t}$.

The probability of $\xi_{K,n}^{CH}(\delta)$ is higher than or equal to $1 - 4nK\delta$. If $n \ge 5K$, the number of pulls $T_{k,n}$ by the CH-AS algorithm launched with parameter δ satisfies on $\xi_{K,n}^{CH}(\delta)$

$$-\lambda_k \left(\frac{12\sqrt{n\log(1/\delta)}}{\sum \lambda_{\min}^{3/2}} + 4K \right) \le T_{k,n} - T_{k,n}^* \le \frac{12\sqrt{n\log(1/\delta)}}{\sum \lambda_{\min}^{3/2}} + 4K, \tag{6}$$

for any arm $1 \le k \le K$.

Proof. The proof is reported in Appendix A.2.

We now show how the bound on the number of pulls translates into a regret bound for the CH-AS algorithm.

Theorem 1. Assume that the distributions $\{\nu_k\}_{k=1}^K$ are supported on [0,1]. If the fixed (known in advance) budget is such that $n \geq 5K$, the regret of \mathcal{A}_{CH} , when it runs with the parameter $\delta = n^{-5/2}$, is bounded as

$$R_n(\mathcal{A}_{CH}) \le \frac{39\sqrt{\log(n)}}{n^{3/2}\lambda_{\min}^{5/2}} + \frac{2.9 \times 10^3}{n^2} \frac{(\log n)^{3/2}}{\lambda_{\min}^{11/2}} \left(1 + \frac{1}{\Sigma^{5/2}}\right). \tag{7}$$

Proof. The proof is reported in Appendix A.3. It is mainly based on the last lemma and on the following inequality (Equation A.13):

$$\mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\}\Big] \le \sup_{\xi} \Big(\frac{\sigma_k^2}{T_{k,n}^2}\Big) \mathbb{E}[T_{k,n}].$$

Remark 1. As discussed in Section 2, our objective is to design a sampling strategy capable of estimating the mean values of the arms almost as accurately as the estimations by the optimal allocation strategy, which assumes that the variances of the arms are known. In fact, Theorem 1 shows that the CH-AS algorithm provides a uniformly accurate estimation of the expected values of the arms with a regret $R_n(\mathcal{A}_{CH})$ of order $\tilde{O}(n^{-3/2})$. This regret rate is the same as the one for the GAFS-MAX algorithm in Antos et al. [1]. Note also that this algorithm is efficient for a fixed horizon n, although it might be possible to change it so that it is efficient for any horizon.

Remark 2. The bound displays an inverse dependency on the smallest optimal allocation proportion λ_{\min} . As a result, the bound scales poorly when an arm has a very small variance relative to the others, i.e., $\sigma_k \ll \Sigma$. Note that GAFS-MAX (see [1]) has also a similar dependency on the inverse of λ_{\min} . Moreover, Theorem 1 holds for a budget $n \geq 5K$, whereas the regret bound of GAFS-MAX in [1] requires a condition $n \geq n_0$, in which n_0 is a constant that scales with $1/\lambda_{\min}$. Finally, note that this UCB type of algorithm (CH-AS) enables a much simpler regret analysis than that of GAFS-MAX.

Remark 3. It is clear from Lemma 1 that the inverse dependency on λ_{\min} appears in the bound on the number of pulls and then is propagated to the regret bound. We however believe that this dependency is not an artifact of the analysis and is intrinsic in the performance of the algorithm. Let us consider a two-arm problem with $\sigma_1^2 = 1/4$ and $\sigma_2^2 = 0$. The optimal allocation is $T_{1,n}^* = n - 1$, $T_{2,n}^* = 1$ (only one sample is enough to estimate the mean of the second arm), and $\lambda_{\min} = 0$. In this case, the arguments used in proving Theorem 1 do not hold anymore and the bound itself becomes vacuous. We conjecture that the Chernoff-Hoeffding's bound used in the upper-confidence term forces the CH-AS to pull the arm with zero variance at least $Dn^{2/3}$ times, where D is a positive constant, with high probability, which results in under-pulling the first arm by the same amount. As a result, the corresponding regret would have a rate of $n^{-4/3}$ w.r.t. the budget n. This suggests that when $\lambda_{\min} = 0$ (or very small compared to 1/n) CH-AS is still able to achieve a o(1/n) regret as the budget n increases but with a slower rate w.r.t. to result proved in Theorem 1.

```
Input: parameters c_1, c_2, \delta
Let a = \sqrt{2c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}}n^{1/2}
Initialize: Pull each arm twice
for t = 2K+1, \ldots, n do

Compute B_{q,t} = \frac{1}{T_{q,t-1}} \left( \hat{\sigma}_{q,t-1}^2 + 4a\hat{\sigma}_{q,t-1} \sqrt{\frac{\log(2/\delta)}{T_{q,t-1}}} + 4a^2 \frac{\log(2/\delta)}{T_{q,t-1}} \right) for each arm 1 \le q \le K
Pull an arm k_t \in \arg\max_{1 \le q \le K} B_{q,t}
end for
Output: \hat{\mu}_{q,t} for all the arms 1 \le q \le K
```

Figure 2: The pseudo-code of the B-AS algorithm. The empirical variances $\hat{\sigma}_{k,t}^2$ are computed according to Equation 8.

Finally, we notice that, for $\lambda_{\min} = 0$, GAFS-MAX is more efficient than CH-AS. In fact, it over-pulls the arms with zero-variance only by $O(n^{1/2})$ and has a regret of order $\tilde{O}(n^{-3/2})$. We will further study how the regret of CH-AS changes with n in Section 5.1.

As discussed in the previous remark, the reason for the poor performance in Lemma 1 for small λ_{\min} can be identified in the fact that Chernoff-Hoeffding's inequality is not tight for small-variance random variables. In Section 4, we propose an algorithm based on a tighter inequality for small-variance random variables, and prove that this algorithm under-pulls all the arms by at most $\tilde{O}(n^{1/2})$, without a dependency on λ_{\min} (see Equations 10 and 11).

4. Allocation Strategy Based on Bernstein UCB

In this section, we present another UCB-like algorithm, called Bernstein Allocation Strategy (B-AS)⁵, based on a tighter variance confidence bound that enables us to improve the bound on $|T_{k,n} - T_{k,n}^*|$ by removing the inverse dependency on λ_{\min} (compare the bounds in Equations 10 and 11 to the one for CH-AS in Equation 6). However this result itself is not sufficient to derive a better regret bound than CH-AS. This finding is interesting since it shows that even an adaptive algorithm which implements a strategy close to the optimal allocation strategy may still incur a regret that poorly scales with the smallest proportion λ_{\min} . We further investigate this issue by showing that the way the bound on the number of pulls translates into a regret bound depends on the specific distributions of the arms. In fact, when the distributions of the arms are Gaussian, we can exploit the property that the empirical variance $\hat{\sigma}_{k,t}^2$ is independent of the empirical mean $\hat{\mu}_{k,t}$, and show that the regret of B-AS no longer depends on $1/\lambda_{\min}$. The numerical simulations in Section 5 further illustrate how the full shape of the distributions (and not only their first two moments) plays an important role in the regret of adaptive allocation algorithms.

4.1. The B-AS Algorithm

The algorithm is based on the use of a high-probability bound, reported in [13] (a similar bound can be found in [2]), on the variance of each arm. Like in the previous section, the arm sampling strategy is determined by those bounds. The B-AS algorithm, A_B , is described in Figure 2. It requires three parameters as input (see Remark 2 in Subsection 4.3 for a discussion on how to reduce the number of parameters from three to one) c_1 and c_2 , which are related to the shape of the distributions (see Assumption 1), and δ , which defines the *confidence level* of the bound. The amount of exploration of the algorithm can be adapted by

⁵The original Bernstein inequality refines the Chernoff-Hoeffding's inequality by introducing the variance of the random variable in the confidence bound. This inequality has been later adapted to the case where the actual variance is unknown and it can be replaced by an empirical estimate of the variance (see [2]). In [13] a similar result is obtained for the variance, where the confidence bound displays a dependency on the empirical estimate of the variance, thus we refer to this algorithm as Bernstein Allocation Strategy. Furthermore, we notice that the inequality derived in [13] does not follow from a trivial application of Chernoff-Hoeffding, since it provides a concentration inequality for the standard deviation which is not an average of i.i.d. random variables but the square root of an average of squared variables.

properly tuning these parameters. The algorithm is similar to CH-AS except that for each arm, the bound $B_{q,t}$ is computed as

$$B_{q,t} = \frac{1}{T_{q,t-1}} \left(\hat{\sigma}_{q,t-1}^2 + 4a\hat{\sigma}_{q,t-1} \sqrt{\frac{\log(2/\delta)}{T_{q,t-1}}} + 4a^2 \frac{\log(2/\delta)}{T_{q,t-1}} \right),$$

where $a = \sqrt{2c_1 \log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}} n^{1/2}$, and⁶

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

Note that actually $\hat{\mu}_{k,t}$, $\hat{\sigma}_{k,t}$, $B_{k,t}$, k_t , and $T_{k,t}$ depend on the arm index (except for k_t), on the time step $t \leq n$, but also, either in a direct or in an indirect way (through the mechanism of the algorithm) on the budget n, on δ which will be chosen as a function of the budget n, and also on c_1 and c_2 . However, since we consider most of the time a fixed budget n and thus a fixed δ , and fixed c_1, c_2 , we conserve this notation in order to have lighter notations.

4.2. Regret Bound and Discussion

The B-AS algorithm is designed to overcome the limitations of CH-AS, especially in the case of arms with different variances. Here we consider a more general assumption than in the previous section, namely that the distributions are sub-Gaussian.

Assumption 1 (Sub-Gaussian distributions). There exist $c_1, c_2 > 0$ such that for all $1 \le k \le K$ and any $\epsilon > 0$,

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

This assumption holds for the Gaussian distribution, and more generally for any distribution whose tail is lighter than Gaussian's. It is thus held for bounded random variables. For example, if $X \in [0,1]$, then the assumption holds with e.g., $c_1 = 1$ and $c_2 = e$.

We first state a bound in Lemma 2 on the difference between the number of pulls suggested by B-AS and the optimal allocation strategy.

Lemma 2. Let Assumption 1 holds for $c_1, c_2 \ge 1$ and let $0 < \delta \le 2/e$. Define the event

$$\xi_{K,n}^{B}(\delta) = \bigcap_{\substack{1 \le k \le K \\ 2 < t < n}} \left\{ \left| \sqrt{\frac{1}{t-1} \sum_{i=1}^{t} \left(X_{k,i} - \frac{1}{t} \sum_{j=1}^{t} X_{k,j} \right)^{2}} - \sigma_{k} \right| \le 2a \sqrt{\frac{\log(2/\delta)}{t}} \right\},\,$$

where $a = \sqrt{2c_1\log(c_2/\delta)} + \frac{\sqrt{c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}}n^{1/2}$. The probability of $\xi_{K,n}^B(\delta)$ is higher than $1 - 2nK\delta$.

When we run the B-AS algorithm with parameters $c_1 \geq 1$, $c_2 \geq 1$, and δ , and budget $n \geq 5K$, on $\xi_{K,n}^B(\delta)$ and for each arm $1 \leq k \leq K$, we have

$$T_{k,n} \ge T_{k,n}^* - K\lambda_k \left[\frac{16a\sqrt{\log(2/\delta)}}{\Sigma} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)} \right) n^{1/2} + 64\sqrt{2K}a^2 \frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}} n^{1/4} + 2 \right], \tag{10}$$

and

$$T_{k,n} \le T_{k,n}^* + K \left[\frac{16a\sqrt{\log(2/\delta)}}{\Sigma} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)} \right) n^{1/2} + 64\sqrt{2K}a^2 \frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}} n^{1/4} + 2 \right], \tag{11}$$

where $c(\delta) = \frac{a\sqrt{3\log(2/\delta)}}{\sqrt{K}(\sqrt{\Sigma} + 3a\sqrt{\log(2/\delta)})}$

Proof. The proof is reported in Appendix B.1 and Appendix B.2.

⁶Unlike in Equation 5, here we use the unbiased estimator of variance.

Remark. Unlike the bounds for CH-AS in Lemma 1, B-AS allocates the pulls on the arms so that, on the event $\xi_{K,n}^B(\delta)$, the bound on the difference between $T_{k,n}$ and $T_{k,n}^*$ is now independent from λ_{\min} , while it preserves a \sqrt{n} dependency on the budget. In practice, this difference may correspond to a significant improvement. In fact, for any finite budget n, if the arms are such that the term depending on λ_{\min} becomes the leading term in the bound in Lemma 1, then we can expect B-AS to outperform CH-AS (see also Remark 3 of Section 3.2 for further discussion of the performance of CH-AS for very small λ_{\min}). Another interesting aspect of the previous lemma is that the lower bound in Equation 10 can be written as $C\lambda_k\sqrt{n}$ (where C>0 does not depend on λ_k). This implies that as allocation ratio λ_k decreases (i.e., arm k should not be pulled much), the difference between $T_{k,n}$ and $T_{k,n}^*$ decreases as well. This is not the case in the upper bound, where the difference between $T_{k,n}$ and $T_{k,n}^*$ does not have any linear dependency on λ_k . This asymmetry between lower and upper bound is the main reason why the final regret bound of B-AS actually displays an inverse dependency on λ_{\min} as shown in Theorem 2.

Theorem 2. Assume that all the distributions $\{\nu_k\}_{k=1}^K$ are sub-Gaussians with parameters c_1 and c_2 . If the fixed (known in advance) budget is such that $n \geq 5K$, the regret of A_B , when it runs with parameters $c_1 \geq 1$, $c_2 \geq 1$, and $\delta = n^{-7/2}$ is bounded as

$$R_n(\mathcal{A}_B) \le \frac{76400c_1(c_2+1)K^2(\log n)^2}{\lambda_{\min}n^{3/2}} + O\left(\frac{(\log n)^6K^7}{n^{7/4}\lambda_{\min}}\right).$$

Proof. The proof is reported in Appendix B.3.

Note again that this algorithm is efficient for a fixed horizon n, although it might be possible to change it so that it is efficient for any horizon.

Similar to Theorem 1, the bound on the number of pulls translates into a regret bound through Equation A.13 reported in Appendix A.3. Note that in order to remove the dependency on λ_{\min} , a symmetric bound on $|T_{k,n} - T_{k,n}^*| \leq \lambda_k \tilde{O}(\sqrt{n})$ is needed. While the lower bound in Equation 10 already decreases with λ_k , the upper bound scales with $\tilde{O}(\sqrt{n})$. Whether there exists an algorithm with a tighter upper bound scaling with λ_k is still an open question. Nonetheless, in the next section, we show that an improved bound on the loss can be achieved in the special case of Gaussian distributions, which leads to a regret bound without the dependency on λ_{\min} .

4.3. Regret for Gaussian Distributions

In the case of Gaussian distributions, the bound on the loss of Equation A.13 can be improved using the following lemma.

Lemma 3. Let $k \leq K$. Assume that the distribution ν_k is Gaussian (and independent of all other distributions $(\nu_{k'})_{k'\neq k}$). Then the loss for arm k of algorithms CH-AS or B-AS satisfies

$$L_{k,n} = \mathbb{E}\left[(\hat{\mu}_{k,n} - \mu_k)^2\right] = \sigma_k^2 \mathbb{E}\left[\frac{1}{T_{k,n}}\right]. \tag{12}$$

Proof. The proof is reported in Appendix C.

Remark. Note that the loss in Equation 12 does not require any upper bound on $T_{k,n}$. It is actually similar to the case of deterministic allocation. When $\tilde{T}_{k,n}$ is the deterministic number of pulls, the corresponding loss resulting from pulling arm k, $\tilde{T}_{k,n}$ times, is $L_{k,n} = \sigma_k^2/\tilde{T}_{k,n}$. In general, when $T_{k,n}$ is a random variable depending on the empirical variances $\{\hat{\sigma}_k^2\}_{k=1}^K$ (like in our adaptive algorithms CH-AS and B-AS), we have

$$\mathbb{E}[(\hat{\mu}_{k,n} - \mu_k)^2] = \sum_{t=1}^n \mathbb{E}[(\hat{\mu}_{k,n} - \mu_k)^2 | T_{k,n} = t] \mathbb{P}[T_{k,n} = t],$$

which might be different than $\sigma_k^2 \mathbb{E}\left[\frac{1}{T_{k,n}}\right]$. In fact, the empirical average $\hat{\mu}_{k,n}$ depends on $T_{k,n}$ through $\{\hat{\sigma}_{k,n}\}_{k=1}^K$, and $\mathbb{E}\left[(\hat{\mu}_{k,n}-\mu_k)^2|T_{k,n}=t\right]$ might not be equal to σ_k^2/t . However, Gaussian distributions have

the property that for any fixed-size sample, the empirical mean is independent from the empirical variance and this enables us to prove Lemma 3, which holds for both the CH-AS and the B-AS algorithm.

We now report a regret bound in the case of the Gaussian distribution. Note that in this case Assumption 1 holds with $c_1 = 2\Sigma$ and $c_2 = 1.7$

Theorem 3. Assume that all the distributions $\{\nu_k\}_{k=1}^K$ are Gaussian and that an upper-bound $\overline{\Sigma} \geq 1/2$ on Σ is known. If the budget is known on advance and such that $n \geq 5K$, the B-AS algorithm launched with parameters $c_1 = 2\overline{\Sigma}$, $c_2 = 1$, and $\delta = n^{-7/2}$ has the following regret bound

$$R_n(\mathcal{A}_B) \le \frac{105 \times 10^3 \bar{\Sigma}}{n^{3/2}} K^2 (\log n)^2$$
 (13)

Proof. The proof is reported in Appendix C.

Remark 1. In the case of Gaussian distributions, the regret bound for B-AS has the rate $\tilde{O}(n^{-3/2})$ without dependency on λ_{\min} , which represents a significant improvement over the regret bounds of the CH-AS and GAFS-MAX algorithms.

Remark 2. In practice, there is no need to tune the three parameters c_1 , c_2 , and δ separately. In fact, it is enough to tune the algorithm for a single parameter $a\sqrt{\log(2/\delta)}$ (see Figure 2). Using the proof of Theorem 2 and the optimized value of δ , as well as the fact that for Gaussian distributions, $c_1 \leq 2\Sigma$, and $c_2 \leq 1$, it is possible to show that choosing a as in Theorem 3 means that $a = O((\overline{\Sigma} \log n)^{1/2})$, where $\overline{\Sigma}$ is an upper bound on the value of Σ . This is a reasonable thing to do whenever a rough estimate of the magnitude of the variances is available.

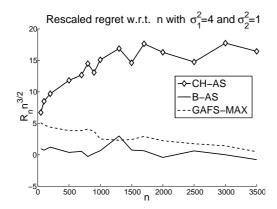
5. Experimental Results

5.1. CH-AS, B-AS, and GAFS-MAX with Gaussian Arms

In this section, we compare the performance of CH-AS, B-AS, and GAFS-MAX on a two-armed problem with Gaussian distributions $\nu_1 = \mathcal{N}(0, \sigma_1^2 = 4)$ and $\nu_2 = \mathcal{N}(0, \sigma_2^2 = 1)$ (note that $\lambda_{\min} = 1/5$). Figure 3-(left) shows the rescaled regret, $n^{3/2}R_n$, for the three algorithms averaged over 50,000 runs. The results indicate that while the rescaled regret is almost constant with respect to n in B-AS and GAFS-MAX, it increases for small (relative to λ_{\min}^{-1}) values of n in CH-AS.

The robust behavior of B-AS when the distributions of the arms are Gaussian may be easily explained by the bound of Theorem 3 (Equation 13). Note though that this experiment seems to imply that there is no additional dependency in $\log(n)$: it could be just an artifact of the proof. The initial increase in the CH-AS curve is also consistent with the bound of Theorem 1 (Equation 7). As discussed in Remark 3 of Section 3.2, we conjecture that the regret bound for CH-AS is of the form $R_n \leq \min \left\{ \lambda_{\min}^{-5/2} \tilde{O}(n^{-3/2}), \tilde{O}(n^{-4/3}) \right\}$, and thus, the algorithm's regret is bounded as $\tilde{O}(n^{-4/3})$ and $\lambda_{\min}^{-5/2} \tilde{O}(n^{-3/2})$ for small and large (relative to λ_{\min}^{-1}) values of n, respectively. It is important to note that the regret bound of CH-AS depends on the arms' distributions only through the variances of the distributions, as shown in Theorem 1. Finally, the curve for GAFS-MAX is very close to the curve for B-AS. For this reason, we believe that it could be possible to improve the GAFS-MAX analysis by using refined concentration inequalities for the standard deviation as done in B-AS. This might also remove the inverse dependency on λ_{\min} and provide a regret bound similar to B-AS in the case of Gaussian distributions.

⁷Note that for a single Gaussian distribution $c_1 = 2\sigma^2$, where σ^2 is the variance of the distribution. Here we use $c_1 = 2\Sigma$ in order for the assumption to be satisfied for all the K distributions simultaneously.



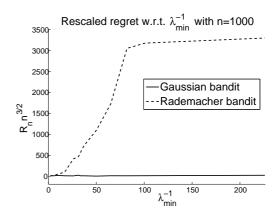


Figure 3: (*left*) The rescaled regret of CH-AS, B-AS, and GAFS-MAX algorithms on a two-armed problem, where the distributions of the arms are Gaussian. (*right*) The rescaled regret of B-AS for two bandit problems, one with two Gaussian arms and one with a Gaussian and a Rademacher arms.

5.2. B-AS with Non-Gaussian Arms

In Section 4.3, we showed that when the arms have Gaussian distribution, the regret bound of the B-AS algorithm no longer depends on λ_{\min} . We also discussed why we conjecture that it is not possible to remove this dependency for general distributions unless a tighter upper bound on the number of pulls can be derived. Although we do not yet have a lower bound on the regret showing the dependency on λ_{\min} , i.e. that the regret might depend on the shape of the distribution, in this section we show that for Rademacher distributions, the regret of B-AS behaves in a different way than for Gaussian distributions with same variance.

As discussed in Section 4.3, the property of the Gaussian distribution that allows us to remove the λ_{\min} dependency in the regret bound of B-AS is that for any sample of fixed size drawn i.i.d. from a Gaussian distribution, the corresponding empirical mean and the empirical variance are independent. The quantities $(\hat{\mu}_{k,n} - \mu_k)^2$ and $\hat{\sigma}_{k,n}$ are however conditionally negatively correlated given $T_{k,n}$ for e.g., the Rademacher distribution.⁸ In the case of Rademacher distribution, the loss $(\hat{\mu}_{k,t} - \mu_k)^2$ is equal to $\hat{\mu}_{k,t}^2$ and we have $\hat{\sigma}_{k,t}^2 = \frac{1}{T_{k,t-1}} \left(\sum_{i=1}^{T_{k,t}} X_{k,i}^2 - T_{k,t} \hat{\mu}_{k,t}^2 \right) = \frac{T_{k,t}}{T_{k,t-1}} \left(1 - \hat{\mu}_{k,t}^2 \right)$, as a result, the larger $\hat{\sigma}_{k,t}^2$ is, the smaller $\hat{\mu}_{k,t}^2$ is. We know that the allocation strategies in CH-AS, B-AS, and GAFS-MAX are based on the empirical variance which is used as a substitute for the true variance. As a result, the larger $\hat{\sigma}_{k,t}^2$ is, the more often arm k is pulled. For the Rademacher distribution, this means that an arm is pulled more than its optimal allocation when its mean is accurately estimated (the loss is small). This may result in a poor estimation of the arm, and thus, negatively affect the regret of the algorithm.

In the experiments of this section, we use B-AS in two different bandit problems: one with two Gaussian arms $\nu_1 = \mathcal{N}(0, \sigma_1^2)$ (with $\sigma_1 \geq 1$) and $\nu_2 = \mathcal{N}(0, 1)$, and one with a Gaussian $\nu_1 = \mathcal{N}(0, \sigma_1^2)$ (with $\sigma_1 \geq 1$) and a Rademacher ν_2 arms. Note that in both cases $\lambda_{\min} = \lambda_2 = 1/(1+\sigma_1^2)$. Figure 3-(right) shows the rescaled regret $(n^{3/2}R_n)$ of the B-AS algorithm as a function of λ_{\min}^{-1} for n=1000. While the rescaled regret of B-AS is constant in the first problem, it increases with σ_1^2 in the second one. This leads us to the conclusion that the shape of the distributions of the arms has an impact on the regret of the algorithm B-AS. In fact, as explained above, this behavior might be due to the poor approximation of the Rademacher arm which is over-pulled exactly whenever its estimated mean is accurate. This result seems to illustrates the fact that in this active learning problem (where the goal is to estimate the mean values of the arms), the performance of the algorithms that rely on the empirical-variance (e.g., CH-AS, B-AS, and GAFS-MAX) depends on the shape of the distributions, and not only on their variances. This may be surprising since according to the central limit theorem the distribution of the empirical mean should tend to a Gaussian. However, it seems that what is important is not the distribution of the empirical mean or variance, but the

 $^{^8}X$ is Rademacher if $X \in \{-1,1\}$ and admits values -1 and 1 with equal probability.

correlation of these two quantities. This is why we believe that any algorithm that is based on empirical standard deviations might be subject to the same problem. However, at the moment no full satisfactory theoretical analysis is available on this point.

6. Conclusions and Open Questions

In this paper, we studied the problem of adaptive allocation for finding a uniformly good estimation of the mean values of K independent distributions. This problem was first studied by Antos et al. [1]. Although the algorithm proposed in [1] achieves a small regret of order $\tilde{O}(n^{-3/2})$, it displays an inverse dependency on the smallest proportion λ_{\min} . In this paper, we first introduced a novel class of algorithms based on upper-confidence-bounds on the (unknown) variances of the arms, and analyzed two such algorithms: Chernoff-Hoeffding allocation strategy (CH-AS) and Bernstein allocation strategy (B-AS). For CH-AS we derived a regret similar to [1], scaling as $\tilde{O}(n^{-3/2})$ and with the dependence on λ_{\min} . Unlike in [1], this result holds for any $n \geq 5K$ and the constants in the bound are made explicit. We then introduced a more refined algorithm, B-AS, whose regret bound does not depend on λ_{\min} for Gaussian arms. Nonetheless, its general regret bound still depends on λ_{\min} . We show that this dependency may be related to the specific distributions of the arms and can be removed for the case of Gaussian distributions. Finally, we report numerical simulations supporting the idea that the shape of the distributions has an impact on the performance of the allocation strategies.

This work opens a number of questions.

- Distribution dependency. Another open question is to which extent the result of B-AS in the case of the Gaussian distribution can be extended to more general families of distributions. As illustrated in the case of Rademacher, the correlation between the empirical mean and variance may cause the algorithm to over-pull arms even when their estimation is accurate, thus incurring a large regret. On the other hand, if the distributions of the arms are Gaussian, their empirical mean and variance are uncorrelated and the allocation algorithms such as B-AS achieve a better regret. Further investigation is needed to identify whether this result can be extended to other distributions.
- Lower bound. The results of Sections 4.3 and 5.2 suggest that the dependency on the distributions of the arms could be intrinsic to the allocation problem. If this is the case, it should be possible to derive a lower bound for this problem showing such dependency (a lower-bound with dependency on λ_{\min}^{-1}). As a matter of fact, no lower bounds are available for this problem and it would be interesting to provide some.

Acknowledgment

This work was supported by French National Research Agency (ANR) through the projects EXPLO-RA n° ANR-08-COSI-004 and LAMPADA n° ANR-09-EMER-007, by Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council and FEDER through the "contrat de projets état region (CPER) 2007–2013", European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231495, and by PASCAL2 European Network of Excellence.

References

- [1] András Antos, Varun Grover, and Csaba Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411:2712–2728, June 2010.
- [2] J-Y. Audibert, R. Munos, and Cs. Szepesvari. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.
- [3] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proceedings of the Twenty-Third Annual Conference on Learning Theory (COLT'10)*, pages 41–53, 2010.
- [4] P. Brémaud. An Introduction to Probabilistic Modeling. Springer, 1988.
- [5] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. Theoretical Computer Science, 412:1832–1852, April 2011. ISSN 0304-3975.

- [6] R. Castro, R. Willett, and R. Nowak. Faster rates in regression via active learning. In Proceedings of Neural Information Processing Systems (NIPS), pages 179–186, 2005.
- [7] P. Chaudhuri and P.A. Mykland. On efficient designing of nonlinear experiments. Statistica Sinica, 5:421-440, 1995.
- [8] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *J. Artif. Int. Res.*, 4:129–145, March 1996. ISSN 1076-9757.
- [9] Morris L Eaton. Multivariate statistics: a vector space approach. Wiley New York, 1983.
- [10] Pierre Étoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12:335–360, 2010.
- [11] V. Fedorov. Theory of Optimal Experiments. Academic Press, 1972.
- [12] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963. URL http://www.jstor.org/stable/2282952?
- [13] A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.

Appendix A. Regret Bound for the CH-AS Algorithm

Let us consider n > 0 and $\delta > 0$ (that can be a function of n) fixed. We consider all the quantities considered in the definition of algorithm CH-AS defined with respect to these fixed n, δ , and use the abbreviated notations $\hat{\mu}_{k,t}$, $\hat{\sigma}_{k,t}$, $B_{k,t}$, k_t , and $T_{k,t}$.

Appendix A.1. Basic Tools

Since the basic tools used in the proof of Theorem 1 are similar to those used in the work by Antos et al. [1], we begin this section by restating two results from that paper. Let ξ be the event

$$\xi = \xi_{K,n}^{CH}(\delta) = \bigcap_{\substack{1 \le k \le K \\ 1 \le t \le n}} \left\{ \left| \left(\frac{1}{t} \sum_{i=1}^{t} X_{k,i}^2 - \left(\frac{1}{t} \sum_{i=1}^{t} X_{k,i} \right)^2 \right) - \sigma_k^2 \right| \le 3\sqrt{\frac{\log(1/\delta)}{2t}} \right\}. \tag{A.1}$$

Note that the first term in the absolute value in Equation (A.1) is the sample variance of arm k computed as in Equation (5) for t samples. It can be shown using Hoeffding's inequality (see Hoeffding [12]) that $\Pr[\xi] \geq 1 - 4nK\delta$, and this is shown by directly reusing the elements of the proof of Lemma 2 in Antos et al. [1]. The event ξ plays an important role in the proofs of this section and several statements will be proved on this event. We now report the following proposition which is analog to Lemma 2 in Antos et al. [1].

Proposition 1. For any k = 1, ..., K and t = 1, ..., n, let $\{X_{k,i}\}_{i=1,...,T_{k,t}}$ be $T_{k,t} \in \{1,...,t\}$ i.i.d. random variables bounded in [0,1] from the distribution ν_k with variance σ_k^2 , and $\hat{\sigma}_{k,t}^2$ be the sample variance computed as in Equation (5). Then the following statement holds on the event ξ :

$$|\hat{\sigma}_{k,t}^2 - \sigma_k^2| \le 3\sqrt{\frac{\log(1/\delta)}{2T_{k,t}}}$$
 (A.2)

We also need to draw a connection between the allocation and stopping time problems. Thus, we report the following proposition which is Lemma 10 in Antos et al. [1].

Proposition 2. 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 with respect to \mathcal{F}_t . Then

$$\mathbb{E}\left[\left(\sum_{i=1}^{T} X_i - T \mu\right)^2\right] = \mathbb{E}[T] \sigma^2. \tag{A.3}$$

Appendix A.2. Allocation Performance

In this subsection, we first provide the proof of Lemma 1 and then use the result in the next subsection to prove Theorem 1.

Proof of Lemma 1. The proof consists of the following three main steps. We assume that ξ holds until the end of this proof.

Step 1. Mechanism of the algorithm. Recall the definition of the upper bound used in A_{CH} at a time t+1>2K:

$$B_{q,t+1} = \frac{1}{T_{q,t}} \left(\hat{\sigma}_{q,t}^2 + 3\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}} \right), \quad 1 \le q \le K.$$

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

$$\frac{\sigma_q^2}{T_{q,t}} \le B_{q,t+1} \le \frac{1}{T_{q,t}} \left(\sigma_q^2 + 6\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}} \right). \tag{A.4}$$

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

$$B_{p,t+1} \le B_{k,t+1} . \tag{A.5}$$

From Equation (A.4) and the fact that $T_{k,t} = T_{k,n} - 1$, we obtain

$$B_{k,t+1} \le \frac{1}{T_{k,t}} \left(\sigma_k^2 + 6\sqrt{\frac{\log(1/\delta)}{2T_{k,t}}} \right) = \frac{1}{T_{k,n} - 1} \left(\sigma_k^2 + 6\sqrt{\frac{\log(1/\delta)}{2(T_{k,n} - 1)}} \right). \tag{A.6}$$

Using the lower bound in Equation (A.4) and the fact that $T_{p,t} \leq T_{p,n}$, we may lower bound $B_{p,t+1}$ as

$$B_{p,t+1} \ge \frac{\sigma_p^2}{T_{n,t}} \ge \frac{\sigma_p^2}{T_{n,r}} \ . \tag{A.7}$$

Combining Equations A.5, A.6, and A.7, we obtain

$$\frac{\sigma_p^2}{T_{p,n}} \le \frac{1}{T_{k,n} - 1} \left(\sigma_k^2 + 6\sqrt{\frac{\log(1/\delta)}{2(T_{k,n} - 1)}} \right). \tag{A.8}$$

Note that at this point there is no dependency on t, and thus, Equation (A.8) holds on the event ξ for any arm k that is pulled at least once after the initialization, and for any arm p.

Step 2. Lower bound on $T_{p,n}$. If an arm q is under-pulled without taking into account the initialization phase, i.e., $T_{q,n}-2<\lambda_q(n-2K)$, then from the constraint $\sum_k(T_{k,n}-2)=n-2K$, we deduce that there must be at least one arm k that is over-pulled, i.e., $T_{k,n}-2>\lambda_k(n-2K)$. Note that for this arm, $T_{k,n}-2>\lambda_k(n-2K)\geq 0$, so we know that this specific arm is pulled at least once after the initialization phase and that it satisfies Equation (A.8). Using the definition of the optimal (up to rounding effects) allocation $T_{k,n}^*=n\lambda_k=n\sigma_k^2/\Sigma$ and the fact that $T_{k,n}\geq \lambda_k(n-2K)+2$, Equation (A.8) may be written as

$$\frac{\sigma_p^2}{T_{p,n}} \le \frac{1}{T_{k,n}^*} \frac{n}{n - 2K} \left(\sigma_k^2 + 6\sqrt{\frac{\log(1/\delta)}{2(\lambda_k(n - 2K) + 2 - 1)}} \right)
\le \frac{\Sigma}{n - 2K} + \frac{12\sqrt{\log(1/\delta)}}{(\lambda_{\min}n)^{3/2}}
\le \frac{\Sigma}{n} + \frac{12\sqrt{\log(1/\delta)}}{(\lambda_{\min}n)^{3/2}} + \frac{4K\Sigma}{n^2},$$
(A.9)

since $\lambda_k(n-2K)+1 \geq \lambda_k(n/2-2K+2K)+1 \geq \frac{n\lambda_k}{2}$, as $n \geq 5K$ (thus also $\frac{2K\Sigma}{n(n-2K)} \leq \frac{4K\Sigma}{n^2}$). Also, if no arm is under-pulled after time 2K, then for each p, $T_{p,n} \geq 2 + \lambda_p(n-2K) > \lambda_p(n-2K)$, i.e., $\sigma_p^2/T_{p,n} \leq \sigma_p^2/(\lambda_p(n-2K)) = \Sigma/(n-2K)$, i.e., Equation (A.9) holds anyway (whether there are underpulled arms or not). By reordering the terms in the previous equation, we obtain the lower bound

$$T_{p,n} \ge \frac{\sigma_p^2}{\frac{\sum}{n} + \frac{12\sqrt{\log(1/\delta)}}{(n\lambda_{\min})^{3/2}} + \frac{4K\Sigma}{n^2}} \ge T_{p,n}^* - \lambda_p \frac{12}{\sum \lambda_{\min}^{3/2}} \sqrt{n\log(1/\delta)} - 4\lambda_p K, \tag{A.10}$$

where in the second inequality we used $1/(1+x) \ge 1-x$ (for x > -1). Note that the lower bound A.10 holds on ξ for any arm p.

Step 3. Upper bound on $T_{p,n}$. Using Equation (A.10) and the fact that $\sum_k T_{k,n} = \sum_k T_{k,n}^* = n$, we obtain the upper bound

$$T_{p,n} = n - \sum_{k \neq p} T_{k,n} \le T_{p,n}^* + \frac{12}{\sum \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} + 4K$$
 (A.11)

The claim follows by combining the lower and upper bounds in Equations A.10 and A.11.

Appendix A.3. Regret Bound

We now show how the bound on the allocation over arms translates into a bound on the regret of the algorithm as stated in Theorem 1.

Proof of Theorem 1. The proof consists of the following two main steps.

Step 1. For each $1 \leq n' \leq n$, $T_{k,n'}$ is a stopping time. 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"). Let $\mathcal{G}_t^{(k)} = \mathcal{G}_t^{(k,n)} = \sigma(\mathcal{F}_t^{(k)}, \mathcal{E}_{-k})$.

We prove for fixed budget n by induction for $n' = 1, \ldots, n$ that each $T_{k,n'}$ is a stopping time with respect

to the filtration $(\mathcal{G}_t^{(k)})_{t \leq n}$.

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 $2K \le n' < n$, and for any t, $\{T_{k,n'} \le t\}$ 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

$$\{T_{k n'+1} \le t\} = (\{T_{k n'} = t\} \cap \{k_{n'+1} \ne k\}) \cup \{T_{k n'} \le t - 1\}.$$
 (A.12)

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 (A.12), 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 with respect to the filtration $(\mathcal{G}_t^{(k)})_{t\leq n}$. **Step 2. Regret bound.** Using its definition, we may write $L_{k,n}$ as follow:

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

Using the definition of $\hat{\mu}_{k,n}$ and Proposition 2 for filtration $\{\mathcal{G}_t^{(k)}\}_{t\leq n}$, $\{X_{k,t}\}_{t\leq n}$, and $T_{k,n}$ (and that $\mathcal{G}_t^{(k)} = \sigma(\{X_{k,t'}: t'\leq t\} \cup \{X_{k',t'}: t'\leq n, k'\neq k\})$ and $\sigma(\{X_{k,t'}: t'\geq t+1\})$ are independent for any $t\leq n$) we

bound the first term as

$$\mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\}\Big] \leq \sup_{\omega \in \xi} \left(\frac{\sigma_k^2}{T_{k,n}^2(\omega)}\right) \mathbb{E}\Big[\frac{\left(\sum_{t=1}^{T_{k,n}} X_{k,t} - T_{k,n} \mu_k\right)^2}{\sigma_k^2} \mathbb{I}\{\xi\}\Big]
\leq \sup_{\xi} \left(\frac{\sigma_k^2}{T_{k,n}^2}\right) \mathbb{E}\Big[\frac{1}{\sigma_k^2} \left(\sum_{t=1}^{T_{k,n}} X_{k,t} - T_{k,n} \mu_k\right)^2\Big]
= \sup_{\xi} \left(\frac{\sigma_k^2}{T_{k,n}^2}\right) \frac{1}{\sigma_k^2} \sigma_k^2 \mathbb{E}[T_{k,n}]
= \sup_{\xi} \left(\frac{\sigma_k^2}{T_{k,n}^2}\right) \mathbb{E}[T_{k,n}],$$
(A.13)

Since the upper-bound in Lemma 1 is obtained on the event ξ (and thus with high probability), and as $T_{k,n} \leq n$, we may easily convert it to a bound in expectation as follows:

$$\mathbb{E}[T_{k,n}] \le \left(T_{k,n}^* + \frac{12}{\sum \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} + 4K\right) + n \times 4nK\delta. \tag{A.14}$$

Combining Equation (A.13) and A.14, and using Equation (A.9) for $\sup_{\xi} \left(\sigma_k^2 / T_{k,n} \right)$, we obtain

$$\mathbb{E}\left[\left(\hat{\mu}_{k,n} - \mu_{k}\right)^{2} \mathbb{I}\left\{\xi\right\}\right] \\
\leq \left(\frac{\Sigma}{n} + \frac{12\sqrt{\log(1/\delta)}}{(\lambda_{\min}n)^{3/2}} + \frac{4K\Sigma}{n^{2}}\right)^{2} \frac{\left(T_{k,n}^{*} + \frac{12}{\Sigma\lambda_{\min}^{3/2}}\sqrt{n\log(1/\delta)} + 4K + n \times 4nK\delta\right)}{\sigma_{k}^{2}}.$$
(A.15)

By setting $A = \frac{12\sqrt{\log(1/\delta)}}{\lambda_{\min}^{3/2}}$ to simplify the notation, Equation (A.15) may be simplified as

$$\begin{split} &\mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\} \Big] \\ & \leq \left(\frac{\Sigma}{n} + \frac{A}{n^{3/2}} + \frac{4K\Sigma}{n^2} \right)^2 \left(\frac{n}{\Sigma} + \frac{A}{\Sigma \sigma_k^2} \sqrt{n} + \frac{4K + 4n^2 K \delta}{\sigma_k^2} \right) \\ & = \left(\frac{\Sigma^2}{n^2} + \frac{A^2}{n^3} + \frac{16K^2 \Sigma^2}{n^4} + \frac{2A\Sigma}{n^{5/2}} + \frac{8K\Sigma^2}{n^3} + \frac{8AK\Sigma}{n^{7/2}} \right) (\cdots) \\ & = \left(\frac{\Sigma^2}{n^2} + \frac{2A\Sigma}{n^{5/2}} + \frac{1}{n^3} \left(A^2 + \frac{16K^2 \Sigma^2}{n} + 8K\Sigma^2 + \frac{8AK\Sigma}{n^{1/2}} \right) \right) (\cdots), \\ & \leq \left(\frac{\Sigma^2}{n^2} + \frac{2A\Sigma}{n^{5/2}} + \frac{1}{n^3} \left(A^2 + 12K\Sigma^2 + 4A\sqrt{K}\Sigma \right) \right) (\cdots), \end{split}$$

where in the last passage we used $n \geq 5K$. Let $B = A^2 + 12K\Sigma^2 + 4A\sqrt{K}\Sigma$. We further simplify the previous expression as

$$\begin{split} & \mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\} \Big] \\ & \leq \frac{\Sigma}{n} + \frac{1}{n^{3/2}} \Big(\frac{\Sigma A}{\sigma_k^2} + 2A \Big) + \frac{1}{n^2} \Big(\frac{4K\Sigma^2}{\sigma_k^2} + \frac{2A^2}{\sigma_k^2} + \frac{B}{\Sigma} \Big) + \frac{1}{n^{5/2}} \Big(\frac{8\Sigma AK}{\sigma_k^2} + \frac{AB}{\sigma_k^2 \Sigma} \Big) + \frac{4KB}{\sigma_k^2 n^3} \\ & + \Big(\frac{4K\Sigma^2}{\sigma_k^2} + \frac{8\Sigma AK}{\sigma_k^2 n^{1/2}} + \frac{4KB}{\sigma_k^2 n} \Big) \delta. \end{split}$$

We now choose $\delta = n^{-5/2}$ and by using $n \ge 5K$ and $\lambda_{\min} \le 1/K$ we obtain

$$\begin{split} &\mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\} \Big] \\ &\leq \frac{\Sigma}{n} + \frac{1}{n^{3/2}} \Big(\frac{\Sigma A}{\sigma_k^2} + 2A \Big) + \frac{1}{n^2} \Big(\frac{4K\Sigma^2}{\sigma_k^2} + \frac{2A^2}{\sigma_k^2} + \frac{B}{\Sigma} + \frac{4\Sigma A\sqrt{K}}{\sigma_k^2} + \frac{AB}{2\sqrt{K}\sigma_k^2\Sigma} + \frac{B}{\sigma_k^2} + \frac{2\Sigma^2\sqrt{K}}{\sigma_k^2} + \frac{B}{2\sqrt{K}\sigma_k^2} \Big) \\ &\leq \frac{\Sigma}{n} + \frac{1}{n^{3/2}} \Big(\frac{\Sigma A}{\sigma_k^2} + 2A \Big) + \frac{1}{\lambda_{\min}n^2} \Big(4K\Sigma + \frac{2A^2}{\Sigma} + \frac{B}{K\Sigma} + 4A\sqrt{K} + \frac{AB}{2\Sigma^2\sqrt{K}} + \frac{B}{\Sigma} + 2\Sigma\sqrt{K} + 2A + \frac{B}{2\sqrt{K}\Sigma} \Big) \\ &= \frac{\Sigma}{n} + \frac{1}{n^{3/2}} \Big(\frac{\Sigma A}{\sigma_k^2} + 2A \Big) + \frac{1}{\lambda_{\min}n^2} \Big(4K\Sigma + 2\Sigma\sqrt{K} + 4A\sqrt{K} + 2A + \frac{2A^2}{\Sigma} + \frac{B}{\Sigma} + \frac{B}{2\sqrt{K}\Sigma} + \frac{B}{K\Sigma} + \frac{AB}{2\Sigma^2\sqrt{K}} \Big) \\ &\leq \frac{\Sigma}{n} + \frac{1}{n^{3/2}} \Big(\frac{\Sigma A}{\sigma_k^2} + 2A \Big) + \frac{1}{\lambda_{\min}n^2} \Big(1.4K^2 + A(4\sqrt{K} + 2) + \frac{2A^2}{\Sigma} + \frac{B}{\Sigma} + \frac{B}{4\Sigma^{3/2}} + \frac{AB}{K\Sigma} + \frac{AB}{4\Sigma^{5/2}} \Big), \end{split}$$

where the last passage follows from $\Sigma \leq K/4$.

Before proceeding further we notice that $\lambda_{\min} \leq 1/K$ and thus

$$K^{3/2} \le \frac{1}{\lambda_{\min}^{3/2}} = \frac{A}{12\sqrt{\log(1/\delta)}} \le \frac{A}{12\sqrt{(5/2)\log n}} \le \frac{A}{27}$$

where the first passage follows from the definition of A and the second from $\delta = n^{-5/2}$, and $n \ge 5K \ge 10$. This implies by definition of B

$$B = A^2 + 12K\Sigma^2 + 4A\sqrt{K}\Sigma \le A^2 + 3A^2/27^2/4 + A^2/27 = 1009A^2/972 < 27A^2/26 < 1.05A^2/27 = 1009A^2/972 < 10$$

where we use $\Sigma \leq K/4$. By using the previous bound, we finally obtain since $1.4K^2 \leq 0.7K^3 \leq 0.7A^2/27^2 \leq 0.7K^3 \leq$ $A^2/1041$

$$\begin{split} &\mathbb{E}\left[(\hat{\mu}_{k,n} - \mu_{k})^{2}\mathbb{I}\{\xi\}\right] \\ &\leq \frac{\Sigma}{n} + \frac{1}{n^{3/2}}\Big(\frac{\Sigma A}{\sigma_{k}^{2}} + 2A\Big) + \frac{1}{\lambda_{\min}n^{2}}\Big(1.4K^{2} + A(4\sqrt{K} + 2) + \frac{2A^{2}}{\Sigma} + \frac{1.05A^{2}}{\Sigma} + \frac{1.05A^{2}}{4\Sigma^{3/2}} + \frac{1.05A^{2}}{K\Sigma} + \frac{1.05A^{3}}{4\Sigma^{5/2}}\Big) \\ &\leq \frac{\Sigma}{n} + \frac{1}{n^{3/2}}\Big(\frac{\Sigma A}{\sigma_{k}^{2}} + 2A\Big) + \frac{1}{\lambda_{\min}n^{2}}\Big(A^{2}/1041 + 0.9A^{3/2} + 3.6\Big(\frac{1}{\Sigma} + \frac{1}{\Sigma^{2}}\Big)A^{2} + \frac{1.05A^{3}}{4\Sigma^{5/2}}\Big) \\ &\leq \frac{\Sigma}{n} + \frac{1}{n^{3/2}}\frac{2A}{\lambda_{\min}n^{2}} + \frac{1}{\lambda_{\min}n^{2}}\Big(0.9A^{3/2} + 3.7\Big(\frac{1}{\Sigma} + \frac{1}{\Sigma^{2}}\Big)A^{2} + \frac{0.27A^{3}}{\Sigma^{5/2}}\Big). \end{split}$$

Since $|\hat{\mu}_{k,n} - \underline{\mu_k}|$ is always smaller than 1, we have $\mathbb{E}[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi^C\}] \le 4nK\delta = 4Kn^{-3/2}$. We also know that $A \leq \frac{19\sqrt{\log(n)}}{\lambda^{3/2}}$. Thus the expected loss of arm k is bounded by

$$L_{k,n} \leq \frac{\Sigma}{n} + \frac{38\sqrt{\log(n)}}{n^{3/2}\lambda_{\min}^{5/2}} + \frac{1}{\lambda_{\min}n^2} \left(0.9A^{3/2} + 3.7\left(\frac{1}{\Sigma} + \frac{1}{\Sigma^2}\right)A^2 + \frac{0.27A^3}{\Sigma^{5/2}}\right) + 4nK\delta$$

$$\leq \frac{\Sigma}{n} + \frac{39\sqrt{\log(n)}}{n^{3/2}\lambda_{\min}^{5/2}} + \frac{2.9 \times 10^3}{n^2} \frac{(\log n)^{3/2}}{\lambda_{\min}^{11/2}} \left(1 + \frac{1}{\Sigma^{5/2}}\right),$$

since $\frac{1}{\Sigma^2} \leq \frac{1}{5} + \frac{4}{5\Sigma^{5/2}}$. Using the definition of regret $R_n(\mathcal{A}) = \max_k L_{k,n} - \frac{\Sigma}{n}$, we obtain

$$R_n(\mathcal{A}_{CH}) \le \frac{39\sqrt{\log(n)}}{n^{3/2}\lambda_{\min}^{5/2}} + \frac{2.9 \times 10^3}{n^2} \frac{(\log n)^{3/2}}{\lambda_{\min}^{11/2}} \left(1 + \frac{1}{\Sigma} + \frac{1}{\Sigma^2} + \frac{1}{\Sigma^{5/2}}\right). \tag{A.16}$$

Appendix B. Regret Bound for the Bernstein Algorithm

Let us consider n > 0, $0 < \delta < 1$ (that can be a function of n), $c_1 > 0$ and $c_2 > 0$ fixed. We consider all the quantities considered in the definition of algorithm B-AS defined with respect to these fixed n, δ, c_1, c_2 , and use the abbreviated notations $\hat{\mu}_{k,t}$, $\hat{\sigma}_{k,t}$, $B_{k,t}$, k_t , and $T_{k,t}$.

Appendix B.1. Basic Tools

Before proving the bound in Theorems 2 and 3 we need a number of technical tools, in particular for sub-Gaussian random variables.

The upper confidence bounds $B_{k,t}$ used in the B-AS algorithm is motivated by Theorem 10 in [13]. We extend this result to sub-Gaussian random variables. We first restate Theorem 10 of [13]:

Theorem 4 ([13]). Let X_1, \ldots, X_t be $t \geq 2$ i.i.d. random variables with variance σ^2 and mean μ and such that $\{X_i\}_{i=1}^t \in [0,b]$. Then with probability at least $1-\delta$, we have

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

We now state and prove the following lemma (first statement of Lemma 2).

Lemma 4. Let Assumption 1 holds, and $n \ge 2$, $c_1 > 0$, $c_2 > 0$, and $0 < \delta < \min(1, c_2)$. For the event

$$\xi = \xi_{K,n}^{B}(\delta) = \bigcap_{\substack{1 \le k \le K \\ 2 \le t \le n}} \left\{ \left| \sqrt{\frac{1}{t-1} \sum_{i=1}^{t} \left(X_{k,i} - \frac{1}{t} \sum_{j=1}^{t} X_{k,j} \right)^{2}} - \sigma_{k} \right| \le 2a \sqrt{\frac{\log(2/\delta)}{t}} \right\}, \tag{B.1}$$

where
$$a = 2\sqrt{c_1 \log(c_2/\delta)} + \frac{\sqrt{c_1 \delta(1 + c_2 + \log(c_2/\delta))}}{(1 - \delta)\sqrt{2 \log(2/\delta)}} n^{1/2}$$
, we have $\Pr[\xi] > 1 - 2nK\delta$.

Note that the first term in the absolute value in Equation B.1 is the empirical standard deviation of arm k computed as in Equation 8 for t samples. The event ξ plays an important role in the proofs of this section and a number of statements will be proved on this event.

Proof. Step 1. Truncating sub-Gaussian variables. We want to characterize the conditional mean and variance of the variables $X_{k,t}$ given that $|X_{k,t} - \mu_k| \leq \sqrt{c_1 \log(c_2/\delta)}$. For any non-negative random variable Y and any $b \geq 0$, $\mathbb{E}[Y\mathbb{I}\{Y > b\}] = \int_b^\infty \mathbb{P}[Y > \epsilon] d\epsilon + b\mathbb{P}[Y > b]$. In order to simplify the notation we introduce the deviation random variable $S_{k,t} = X_{k,t} - \mu_k$. If we take $b = c_1 \log(c_2/\delta)$ and use Assumption 1, we obtain $\mathbb{P}[S_{k,t}^2 > b] \leq \delta$ and

$$\mathbb{E}\Big[S_{k,t}^2 \mathbb{I}\left\{S_{k,t}^2 > b\right\}\Big] = \int_b^\infty \mathbb{P}\big[S_{k,t}^2 > \epsilon\big] d\epsilon + b\mathbb{P}[S_{k,t}^2 > b] \le \int_b^\infty c_2 \exp(-\epsilon/c_1) d\epsilon + bc_2 \exp(-b/c_1)$$
$$= c_1 \delta + c_1 \delta \log(c_2/\delta) = c_1 \delta \left(1 + \log(c_2/\delta)\right).$$

By definition of $S_{k,t}$, we have $\mathbb{E}\left[S_{k,t}^2 \mathbb{I}\left\{S_{k,t}^2 > b\right\}\right] + \mathbb{E}\left[S_{k,t}^2 \mathbb{I}\left\{S_{k,t}^2 \leq b\right\}\right] = \sigma_k^2$, which can be written as

$$\frac{\mathbb{E}\left[S_{k,t}^{2}\mathbb{I}\{S_{k,t}^{2} > b\}\right] - \sigma_{k}^{2}\mathbb{P}\left[S_{k,t}^{2} > b\right]}{\mathbb{P}\left[S_{k,t}^{2} \leq b\right]} = \sigma_{k}^{2} - \frac{\mathbb{E}\left[S_{k,t}^{2}\mathbb{I}\{S_{k,t}^{2} \leq b\}\right]}{\mathbb{P}\left[S_{k,t}^{2} \leq b\right]},\tag{B.2}$$

⁹Let $\tilde{Y} = Y\mathbb{I}\{Y > b\} + b\mathbb{I}\{Y \le b\}$, then $\mathbb{E}[\tilde{Y}] = \int_0^b \mathbb{P}[\tilde{Y} > \varepsilon] d\varepsilon + \int_b^\infty \mathbb{P}[\tilde{Y} > \varepsilon] d\varepsilon = b + \int_b^\infty \mathbb{P}[Y > \varepsilon] d\varepsilon$. Thus we can write $\mathbb{E}[Y\mathbb{I}\{Y > b\}] = \mathbb{E}[\tilde{Y}] - b\mathbb{P}[Y \le b] = \int_b^\infty \mathbb{P}[Y > \varepsilon] d\varepsilon + b\mathbb{P}[Y > b]$.

that combined with the previous equation, implies that

$$\left| \mathbb{E} \left[S_{k,t}^2 \middle| S_{k,t}^2 \le b \right] - \sigma_k^2 \right| = \frac{\left| \mathbb{E} \left[\left(S_{k,t}^2 - \sigma_k^2 \right) \mathbb{I} \left\{ S_{k,t}^2 > b \right\} \right] \right|}{\mathbb{P} \left[S_{k,t}^2 \le b \right]}$$

$$\leq \frac{c_1 \delta (1 + \log(c_2/\delta)) + \delta \sigma_k^2}{1 - \delta}, \tag{B.3}$$

where we use $1 + \log(c_2/\delta) \ge 0$, that follows from $\delta \le c_2$. Note also that Cauchy-Schwartz inequality implies

$$\left| \mathbb{E} \left[S_{k,t} \mathbb{I} \left\{ S_{k,t}^2 > b \right\} \right] \right| \leq \sqrt{\mathbb{E} \left[S_{k,t}^2 \mathbb{I} \left\{ S_{k,t}^2 > b \right\} \right]}$$

$$\leq \sqrt{c_1 \delta (1 + \log(c_2/\delta))}.$$

We now introduce the conditional mean of $X_{k,t}$ conditioned on small deviations, that is $\tilde{\mu}_k = \mathbb{E}\left[X_{k,t} \middle| S_{k,t}^2 \leq b\right] = \frac{\mathbb{E}\left[X_{k,t} \mathbb{E}\left\{S_{k,t}^2 \leq b\right\}\right]}{\mathbb{E}\left[S_{k,t}^2 \leq b\right]}$. Thus we can combine $\mathbb{E}\left[X_{k,t} \mathbb{E}\left\{S_{k,t}^2 > b\right\}\right] + \mathbb{E}\left[X_{k,t} \mathbb{E}\left\{S_{k,t}^2 \leq b\right\}\right] = \mu_k$ with the previous result and obtain

$$|\tilde{\mu}_k - \mu_k| = \frac{\left| \mathbb{E}\left[S_{k,t} \mathbb{I}\left\{ S_{k,t}^2 > b \right\} \right] \right|}{\mathbb{P}\left[S_{k,t}^2 \le b \right]} \le \frac{\sqrt{c_1 \delta (1 + \log(c_2/\delta))}}{1 - \delta}.$$
 (B.4)

We also define the variance of the conditional random variable $\tilde{\sigma}_k^2 = \mathbb{V}[X_{k,t}|S_{k,t}^2 \leq b] = \mathbb{E}[S_{k,t}^2|S_{k,t}^2 \leq b] - (\mu_k - \tilde{\mu_k})^2$. From Equations B.3 and B.4, we derive

$$\begin{split} |\tilde{\sigma}_{k}^{2} - \sigma_{k}^{2}| &\leq \left| \mathbb{E} \left[S_{k,t}^{2} | S_{k,t}^{2} \leq b \right] - \sigma_{k}^{2} \right| + (\tilde{\mu}_{k} - \mu_{k})^{2} \\ &\leq \frac{c_{1}\delta(1 + \log(c_{2}/\delta)) + \delta\sigma_{k}^{2}}{1 - \delta} + \frac{c_{1}\delta(1 + \log(c_{2}/\delta))}{(1 - \delta)^{2}} \\ &\leq \frac{2c_{1}\delta(1 + \log(c_{2}/\delta)) + \delta\sigma_{k}^{2}}{(1 - \delta)^{2}}. \end{split}$$

In order to get the final result, we first bound the variance σ_k^2 as a function of the constants c_1 and c_2 using the sub-Gaussian assumption as

$$\sigma_k^2 = \mathbb{E}[(X_{k,t} - \mu_k)^2] = \int_0^\infty \mathbb{P}[X_{k,t} - \mu_k)^2 > \varepsilon d\varepsilon \le \int_0^\infty c_2 \exp(-\varepsilon/c_1) d\varepsilon = c_1 c_2.$$
 (B.5)

Finally, using $\sqrt{|x^2-y^2|} \ge |x-y|$ for $x,y \ge 0$, we obtain

$$|\tilde{\sigma}_k - \sigma_k| \le \frac{\sqrt{2c_1\delta(1 + c_2 + \log(c_2/\delta))}}{1 - \delta}.$$
(B.6)

Step 2. Application of large deviation inequalities.

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

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

Under Assumption 1, using a union bound, we have that the probability of this event is at least $1 - nK\delta$. On ξ_1 , the $\{X_{k,i}\}_i$, $1 \le k \le K$, $1 \le i \le t$ are t i.i.d. bounded random variables with standard deviation $\tilde{\sigma}_k$. Let $\xi_2 = \xi_{2,K,n}(\delta)$ be the event:

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

Using Theorem 4 and a union bound, we deduce that $\Pr[\xi_1 \cap \xi_2] \ge 1 - 2nK\delta$. Now, from Equation (B.6), we have on $\xi_1 \cap \xi_2$, for all $1 \le k \le K$, $2 \le t \le n$:

$$\left| \sqrt{\frac{1}{t-1} \sum_{i=1}^{t} \left(X_{k,i} - \frac{1}{t} \sum_{j=1}^{t} X_{k,j} \right)^{2} - \sigma_{k}} \right|$$

$$\leq 2\sqrt{c_{1} \log(c_{2}/\delta)} \sqrt{\frac{2 \log(2/\delta)}{t-1}} + \frac{\sqrt{2c_{1}\delta(1 + c_{2} + \log(c_{2}/\delta))}}{1 - \delta}$$

$$\leq 4\sqrt{c_{1} \log(c_{2}/\delta)} \sqrt{\frac{\log(2/\delta)}{t}} + \frac{\sqrt{2c_{1}\delta(1 + c_{2} + \log(c_{2}/\delta))}}{1 - \delta},$$

from which we deduce Lemma 4 (since $\xi_1 \cap \xi_2 \subseteq \xi$ and $2 \le t \le n$).

We transcribe the definition (B.1) of ξ in the last lemma into the following lemma when the number of samples $T_{k,t}$ are random.

Lemma 5. For t = 2K, ..., n, let $T_{k,t}$ be any random variable taking values in $\{2, ..., n\}$. Let $\hat{\sigma}_{k,t}^2$ be the empirical variance computed from Equation (8). Then, on the event ξ , we have:

$$|\hat{\sigma}_{k,t} - \sigma_k| \le 2a\sqrt{\frac{\log(2/\delta)}{T_{k,t}}}, \tag{B.7}$$

where
$$a = 2\sqrt{c_1 \log(c_2/\delta)} + \frac{\sqrt{T_{k,t}c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)\sqrt{2\log(2/\delta)}}$$

Appendix B.2. Allocation Performance

In this section, we first provide the proof of Lemma 2, we then derive the regret bound of Theorem 2 in the general case, and we prove Theorem 3 for Gaussians.

Recall that $n \geq 5K$. This will be useful in the following.

Proof of Lemma 2. Note first that the first part of the claim of the lemma is exactly Lemma 4. The rest of the proof consists of the following five main steps. Until the end of the proof, we assume that ξ holds.

Step 1. Lower bound of order $\Omega(\sqrt{n})$. We first recall for any arm q the definition of $B_{q,t+1}$ used in the B-AS algorithm

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

Using Lemma 5 it follows that on ξ , for any q such that $T_{q,t} \geq 2$,

$$\frac{\sigma_q^2}{T_{q,t}} \le B_{q,t+1} \le \frac{1}{T_{q,t}} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right)^2.$$
 (B.8)

Let q be the index of an arm such that $T_{q,n} \geq \frac{n}{K}$ and $t+1 \leq n$ be the last time that it was pulled, i.e., $T_{q,t} = T_{q,n} - 1$ and $T_{q,t+1} = T_{q,n}$. ¹⁰ From Equation (B.8) and the fact that $T_{q,n} \geq \frac{n}{K} \geq 5$ (see condition on $c(\delta)$, and also the beginning of this section) and $T_{q,t} \geq 3$, we obtain on ξ

¹⁰Note that such an arm always exists for any possible allocation strategy given the constraint $n = \sum_{p} T_{p,n}$.

$$B_{q,t+1} \le \frac{1}{T_{q,t}} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right)^2 \le \frac{4K}{3n} \left(\sqrt{\Sigma} + 4a\sqrt{\frac{\log(2/\delta)}{3}} \right)^2, \tag{B.9}$$

where we also used $T_{q,n} \geq 4$ to bound $T_{q,t}$ in the parenthesis and the fact that $\sigma_q \leq \sqrt{\Sigma}$. Since at time t+1 we assumed that arm q has been chosen then for any other arm p, we have

$$B_{p,t+1} \le B_{q,t+1}.$$
 (B.10)

From the definition of $B_{p,t+1}$, removing all the terms but the last and using the fact that $T_{p,t} \leq T_{p,n}$, we obtain the lower bound

$$B_{p,t+1} \ge 4a^2 \frac{\log(2/\delta)}{T_{n,t}^2} \ge 4a^2 \frac{\log(2/\delta)}{T_{p,n}^2}$$
 (B.11)

Combining Equations B.9–B.11, we obtain

$$4a^2 \frac{\log(2/\delta)}{T_{n,n}^2} \le \frac{4K\left(\sqrt{\Sigma} + 3a\sqrt{\log(2/\delta)}\right)^2}{3n}.$$

Finally, this implies that for any p

$$T_{p,n} \ge \frac{2a\sqrt{\log(2/\delta)}}{\sqrt{\Sigma} + 3a\sqrt{\log(2/\delta)}}\sqrt{\frac{3n}{4K}}.$$
 (B.12)

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

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

thus obtaining $T_{p,n} \ge c(\delta)\sqrt{n}$ on the event ξ for any p.

Step 2. Mechanism of the algorithm. Note that as $n \ge 5K$, there is at least an arm q that is pulled after initialization. Let, for such an arm q, t+1 > 2K be the time when arm q is pulled for the last time, that is $T_{q,t} = T_{q,n} - 1 \ge 2$. Since at time t+1 this arm q is chosen, then for any other arm p, we have

$$B_{p,t+1} \le B_{q,t+1}$$
 (B.13)

From Equation (B.8) and $T_{q,t} = T_{q,n} - 1$, we obtain

$$B_{q,t+1} \le \frac{1}{T_{q,t}} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right)^2 = \frac{1}{T_{q,n} - 1} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,n} - 1}} \right)^2.$$
 (B.14)

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

$$B_{p,t+1} \ge \frac{\sigma_p^2}{T_{p,t}} \ge \frac{\sigma_p^2}{T_{p,n}}.$$
 (B.15)

Combining Equations B.13–B.15, we obtain

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

Summing over all q that are pulled after initialization on both sides, we obtain on ξ for any arm p

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \le \sum_{q|T_{q,n}>2} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}}\right)^2,$$
(B.16)

because the arms that are not pulled after the initialization are only pulled twice (so $\sum_{q|T_{q,n}>2} (T_{q,n}-1) \ge n-2K$).

Step 3. Intermediate lower bound. It is possible to rewrite Equation (B.16), using the fact that $T_{q,n} \ge 2$, as

$$\frac{\sigma_p^2}{T_{p,n}}(n - 2K) \le \sum_{q} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,n} - 1}}\right)^2 \le \sum_{q} \left(\sigma_q + 4a\sqrt{\frac{2\log(2/\delta)}{T_{q,n}}}\right)^2.$$
(B.17)

Plugging Equation (B.12) in Equation (B.17), we have on ξ for any arm p

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \le \sum_q \left(\sigma_q + 4a\sqrt{\frac{2\log(2/\delta)}{T_{q,n}}}\right)^2 \le \left(\sqrt{\Sigma} + 4\sqrt{K}a\sqrt{2\frac{\log(2/\delta)}{c(\delta)\sqrt{n}}}\right)^2,\tag{B.18}$$

because for any sequence $(a_k)_{i=1,...,K} \ge 0$, and any $b \ge 0$, $\sum_k (a_k + b)^2 \le (\sqrt{\sum_k a_k^2} + \sqrt{K}b)^2$ by Cauchy-Schwartz.

Building on this bound we shall recover the desired bound.

Step 4. Final lower bound. We first expand the square in Equation (B.17) using $T_{q,n} \geq 2$ as

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \le \sum_q \sigma_q^2 + 8a\sqrt{2\log(2/\delta)} \sum_q \frac{\sigma_q}{\sqrt{T_{q,n}}} + \sum_q \frac{32a^2\log(2/\delta)}{T_{q,n}}.$$

We now use the bound in Equation (B.18) in the second term of the RHS and the bound in Equation (B.12) to bound $T_{k,n}$ in the last term, thus obtaining

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \le \Sigma + 8a\sqrt{2\log(2/\delta)} \frac{K}{\sqrt{n-2K}} \left(\sqrt{\Sigma} + 4\sqrt{K}a\sqrt{2\frac{\log(2/\delta)}{c(\delta)\sqrt{n}}}\right) + \frac{32Ka^2\log(2/\delta)}{c(\delta)\sqrt{n}}$$

By using again $n \geq 5K$ and some algebra, we get

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \le \Sigma + 16a\sqrt{\log(2/\delta)} \frac{K}{\sqrt{n}} \left(\sqrt{\Sigma} + 4\sqrt{K}a\sqrt{2\frac{\log(2/\delta)}{c(\delta)\sqrt{n}}}\right) + \frac{32Ka^2\log(2/\delta)}{c(\delta)\sqrt{n}} \\
\le \Sigma + 16Ka\sqrt{\frac{\Sigma\log(2/\delta)}{n}} + 64\sqrt{2}K^{3/2}a^2\frac{\log(2/\delta)}{\sqrt{c(\delta)}}n^{-3/4} + \frac{32Ka^2\log(2/\delta)}{c(\delta)\sqrt{n}} \\
= \Sigma + \frac{16Ka\sqrt{\log(2/\delta)}}{\sqrt{n}} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right) + 64\sqrt{2}K^{3/2}a^2\frac{\log(2/\delta)}{\sqrt{c(\delta)}}n^{-3/4}. \tag{B.19}$$

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

$$\begin{split} T_{p,n} &\geq \frac{\sigma_p^2(n-3K)}{\Sigma} \left[1 + \frac{16Ka\sqrt{\log(2/\delta)}}{\Sigma\sqrt{n}} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right) + 64\sqrt{2}K^{3/2}a^2\frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}}n^{-3/4} \right]^{-1} \\ &\geq \frac{\sigma_p^2(n-2K)}{\Sigma} \left[1 - \frac{16Ka\sqrt{\log(2/\delta)}}{\Sigma\sqrt{n}} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right) - 64\sqrt{2}K^{3/2}a^2\frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}}n^{-3/4} \right] \\ &\geq T_{p,n}^* - K\lambda_p \left[\frac{16a\sqrt{\log(2/\delta)}}{\Sigma} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right)n^{1/2} + 64\sqrt{2K}a^2\frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}}n^{1/4} + 2 \right]. \end{split}$$

Note that the above lower bound holds on ξ for any arm p.

Step 5. Upper bound. The upper bound on $T_{p,n}$ follows by using $T_{p,n} = n - \sum_{q \neq p} T_{q,n}$ and the previous lower bound, that is

$$\begin{split} T_{p,n} &\leq n - \sum_{q \neq p} T_{q,n}^* \\ &+ \sum_{q \neq p} K \lambda_q \left[\frac{16a\sqrt{\log(2/\delta)}}{\Sigma} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)} \right) n^{1/2} + 64\sqrt{2K}a^2 \frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}} \; n^{1/4} + 2 \right] \\ &\leq T_{p,n}^* + K \left[\frac{16a\sqrt{\log(2/\delta)}}{\Sigma} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)} \right) n^{1/2} + 64\sqrt{2K}a^2 \frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}} \; n^{1/4} + 2 \right]. \end{split}$$

Appendix B.3. Regret Bounds

With the allocation performance, we now move to the regret bound showing how the number of pulls translates into the losses L_{kn} and the global regret as stated in Theorem 2.

We first state some technical results.

Appendix B.3.1. Bound on the Regret Outside ξ

The next lemma provides a bound for the loss whenever the event ξ does not hold.

Lemma 6. Let Assumption 1 holds. If $2nK\delta < c_2$, then for every arm k, we have 11

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

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

$$\mathbb{P}[(X_{k,t} - \mu_k)^2 \ge \epsilon] \le c_2 \exp(-\epsilon/c_1) ,$$

and thus since $c_2 > 2nK\delta$, we obtain

$$\mathbb{P}\big[(X_{k,t} - \mu_k)^2 \ge c_1 \log(c_2/2nK\delta)\big] \le 2nK\delta.$$

Since $\mathbb{P}[\xi^C] \leq 2nK\delta$, the previous equation implies, using $c_2/(2nK\delta) > 1$

$$\mathbb{E}\left[(X_{k,t} - \mu_k)^2 \mathbb{I}\left\{\xi^C\right\}\right] = \int_0^\infty \mathbb{P}\left[(X_{k,t} - \mu_k)^2 \mathbb{I}\left\{\xi^C\right\} > \epsilon\right] d\epsilon$$

$$\leq \int_{c_1 \log(c_2/2nK\delta)}^\infty c_2 \exp(-\epsilon/c_1) d\epsilon + c_1 \log(c_2/2nK\delta) \mathbb{P}[\xi^C]$$

$$\leq 2c_1 nK\delta (1 + \log(c_2/2nK\delta)).$$

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

¹¹Note that for $\delta = n^{-7/2}$, $n \ge 5K$, and $c_2 \ge 1$, we have $2nK\delta = 2Kn^{-5/2} < c_2$.

Appendix B.3.2. Other Technical Inequalities

At first let us write, for the sake of convenience,

$$B = 16 K a \sqrt{\log(2/\delta)} \left(\sqrt{\Sigma} + \frac{2a \sqrt{\log(2/\delta)}}{c(\delta)} \right) \quad \text{and} \quad C = 64 \sqrt{2} K^{3/2} a^2 \frac{\log(2/\delta)}{\sqrt{c(\delta)}}.$$

Upper and lower bound on a. If $\delta = n^{-7/2}$, with $n \geq 5K \geq 10$ and $c_2 \geq 1$

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

$$\leq \sqrt{14c_1(c_2+1)\log(n)} + \frac{2}{n^{5/4}}\sqrt{c_1(1+c_2)} \leq \sqrt{15c_1(c_2+1)\log(n)}$$

$$\leq 4\sqrt{c_1(c_2+1)\log(n)}.$$

We also have by just keeping the first term, since $c_2 \geq 1$

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

Lower bound on $c(\delta)$ when $\delta = n^{-7/2}$. See Lemma 2 for the definition of $c(\delta)$. Using the fact that the arms have sub-Gaussian distribution we showed in Equation (B.5) that $\sigma_k^2 \leq c_1 c_2$, then we also have $\Sigma \leq K c_1 c_2$. If $\delta = n^{-7/2}$, we obtain by using the previous lower bound on a that

$$\begin{split} c(\delta = n^{-7/2}) &= \frac{a\sqrt{3\log(2/\delta)}}{\sqrt{3K}\left(\sqrt{\Sigma/3} + a\sqrt{3\log(2/\delta)}\right)} = \frac{1}{\sqrt{3K}}\left(1 - \frac{\sqrt{\Sigma/3}}{\sqrt{\Sigma/3} + a\sqrt{\log2/\delta}}\right) \\ &\geq \frac{1}{\sqrt{3K}}\left(1 - \frac{\sqrt{\Sigma/3}}{\sqrt{\Sigma/3} + \sqrt{c_1\log2/\delta}}\right) \geq \frac{1}{\sqrt{3K}}\left(1 - \frac{\sqrt{\Sigma/3}}{\sqrt{\Sigma/3} + \sqrt{c_1}}\right) \geq \frac{1}{\sqrt{K}}\left(\frac{1}{\sqrt{Kc_2} + \sqrt{3}}\right) \end{split}$$

by using $\Sigma \leq Kc_2c_1$ for the last step.

Upper bound on the loss outside ξ when $\delta = n^{-7/2}$. We get from Lemma 6 when $\delta = n^{-7/2}$, when $c_2 \ge 1$ and when $n \ge 5K$ that

$$\mathbb{E}\left[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\left\{\xi^C\right\}\right] \le 2c_1 n^2 K \delta\left(1 + \log\left(\frac{c_2}{2nK\delta}\right)\right) \le 2c_1 K n^{-3/2} \left(1 + (c_2 + 1)\log\left(\frac{n^{5/2}}{2K}\right)\right)$$

$$\le 2c_1 K n^{-3/2} \left(1 + \frac{5}{2}(c_2 + 1)\log(n)\right) \le 7c_1 K (c_2 + 1)\log(n) n^{-3/2}.$$

Upper bound on B for $\delta = n^{-7/2}$. See the proof of Theorem 2 for the definition of B (the notation B we use in this section is for technical purposes and has nothing to do with the B introduced in the proofs for algorithm CH-AS). When $\delta = n^{-7/2}$, when $c_2 \geq 1$ and when $n \geq 5K \geq 10$,

$$B = 16Ka\sqrt{\log(2/\delta)} \left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right)$$

$$\leq 16Ka\sqrt{7/2\log(2n)} \left(\sqrt{\Sigma} + 2\sqrt{K}(\sqrt{\Sigma} + 3a\sqrt{7/2\log(2n)})\right)$$

$$\leq 16Ka\sqrt{7/2\log(2n)} \left(\sqrt{\Sigma} + 2\sqrt{K\Sigma} + 12\sqrt{K}\sqrt{c_1(c_2 + 1)7\log(n)\log(2n)}\right)$$

$$\leq 16Ka\sqrt{7/2\log(2n)} \left(3K\sqrt{c_1c_2} + 45\sqrt{K}\sqrt{c_1(c_2 + 1)\log(n)}\right)$$

$$\leq 32K\sqrt{14c_1(c_2 + 1)\log n\log(2n)} \left(48K\sqrt{c_1(c_2 + 1)\log(n)}\right)$$

$$\leq 8 \times 10^3K^2c_1(c_2 + 1)\log^2(n).$$

Upper bound on C for $\delta = n^{-7/2}$. See the proof of Theorem 2 for the definition of C. When $\delta = n^{-7/2}$, when $c_2 \ge 1$ and when $n \ge 5K \ge 10$,

$$C = 64\sqrt{2}K^{3/2}a^{2}\frac{\log(2/\delta)}{\sqrt{c(\delta)}} = 64\sqrt{2}K^{3/2}\frac{a^{2}\log(2/\delta)}{\sqrt{a}(3\log(2/\delta))^{1/4}}K^{1/4}(\sqrt{\Sigma} + 3a\sqrt{\log(2/\delta)})^{1/2}$$

$$\leq 64\sqrt{2}K^{3/2}a^{3/2}(\log(2/\delta))^{3/4}\frac{1}{3^{1/4}}K^{1/4}(\sqrt{Kc_{1}c_{2}} + 12\sqrt{c_{1}(c_{2}+1)\log n}\sqrt{7\log n})^{1/2}$$

$$\leq 128\sqrt{2}\frac{1}{3^{1/4}}K^{7/4}(2\sqrt{2c_{1}(c_{2}+1)\log n})^{3/2}(7\log n)^{3/4}\sqrt{24}K^{1/4}(c_{1}(c_{2}+1))^{1/4}\sqrt{\log n}$$

$$\leq 14 \times 10^{3}K^{2}c_{1}(c_{2}+1)\log^{2}(n).$$

We are now ready to prove Theorem 2.

Proof of Theorem 2. Equation (B.19) becomes using the constants B, C that we introduced

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \le \Sigma + \frac{B}{\sqrt{n}} + \frac{C}{n^{3/4}}.$$
(B.20)

We also have the upper bound in Lemma 2 which can be rewritten:

$$T_{p,n} \le T_{p,n}^* + \frac{B}{\Sigma} \sqrt{n} + \frac{C}{\Sigma} n^{1/4} + 2K.$$

Note that because this upper bound holds on an event of probability bigger than $1 - 4nK\delta$ and also because $T_{p,n}$ is bounded by n anyways, we can convert the former upper bound in a bound in expectation:

$$\mathbb{E}[T_{p,n}] \le T_{p,n}^* + \frac{B}{\Sigma} \sqrt{n} + \frac{C}{\Sigma} n^{1/4} + 2K + n \times 4nK\delta. \tag{B.21}$$

We recall that the loss of any arm k is decomposed in two parts as follows:

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

By combining the fact that $T_{k,n}$ is again a stopping time with Equations B.20, B.21, and A.3 (as done in Equation (A.13)), and since n - 2K > 0, we obtain for the first part of the loss:

$$\begin{split} &\mathbb{E}[(\hat{\mu}_{k,n} - \mu_{k})^{2}\mathbb{I}\left\{\xi\right\}] \\ &\leq \frac{1}{\sigma_{k}^{2}(n-2K)^{2}} \left(\Sigma + \frac{B}{\sqrt{n}} + \frac{C}{n^{3/4}}\right)^{2} \left(T_{k,n}^{*} + \frac{B}{\Sigma}\sqrt{n} + \frac{C}{\Sigma}n^{1/4} + 2K + 4n^{2}K\delta\right) \\ &\leq \frac{1}{(n-2K)^{2}} \left(\Sigma^{2} + 2\Sigma \left(\frac{B}{\sqrt{n}} + \frac{C}{n^{3/4}}\right) + \frac{(B+C)^{2}}{n}\right) \left(\frac{n}{\Sigma} + \frac{B}{\Sigma^{2}\lambda_{k}}\sqrt{n} + \frac{C}{\Sigma^{2}\lambda_{k}}n^{1/4} + \frac{2K}{\Sigma\lambda_{k}} + \frac{4n^{2}K\delta}{\Sigma\lambda_{k}}\right) \\ &\leq \frac{1}{(n-2K)^{2}} \left(n\Sigma + \frac{B}{\lambda_{k}}\sqrt{n} + \frac{C+2K\Sigma}{\lambda_{k}}n^{1/4} + \frac{4n^{2}K\Sigma\delta}{\lambda_{k}} + 2B\sqrt{n} + 2Cn^{1/4} \right. \\ &\quad + \frac{2(B+C)(\frac{B}{\Sigma} + \frac{C}{\Sigma} + 2K)}{\lambda_{k}} + \frac{8(B+C)n^{3/2}K\delta}{\lambda_{k}} + (B+C)^{2} \left(\frac{1}{\Sigma} + \frac{B+C}{\Sigma^{2}\lambda_{k}} + \frac{2K}{\Sigma\lambda_{k}}\right) + 4nK\delta\frac{(B+C)^{2}}{\Sigma\lambda_{k}} \right) \\ &= \frac{1}{(n-2K)^{2}} \left(n\Sigma + \left(\frac{B}{\lambda_{k}} + 2B\right)\sqrt{n} + \left(\frac{C+2K\Sigma}{\lambda_{k}} + 2C\right)n^{1/4} \right. \\ &\quad + \frac{2(B+C)(\frac{B+C}{\Sigma} + 2K)}{\lambda_{k}} + (B+C)^{2} \left(\frac{1}{\Sigma} + \frac{B+C}{\Sigma^{2}\lambda_{k}} + \frac{2K}{\Sigma\lambda_{k}}\right) \right. \\ &\quad + \frac{4n^{2}K\Sigma\delta}{\lambda_{k}} + \frac{8(B+C)n^{3/2}K\delta}{\lambda_{k}} + 4nK\delta\frac{(B+C)^{2}}{\Sigma\lambda_{k}} \right) \\ &\leq \frac{1}{(n-2K)^{2}} \left(n\Sigma + \frac{3B}{\lambda_{k}}\sqrt{n} + \frac{3C+2K\Sigma}{\lambda_{k}}n^{1/4}} \right. \\ &\quad + \frac{K(B+C)^{3}}{\lambda_{k}} \left(\frac{2}{K\Sigma(B+C)} + \frac{4}{(B+C)^{2}} + \frac{\lambda_{k}}{K\Sigma(B+C)} + \frac{1}{\Sigma^{2}K} + \frac{2}{\Sigma(B+C)}\right) \\ &\quad + \frac{4\delta n^{2}K}{\lambda_{k}} \left(\Sigma + 2(B+C) + \frac{(B+C)^{2}}{\Sigma}\right)\right), \end{split}$$

and since $B+C\geq 2$ for $\delta=n^{-7/2},\,n\geq 16K/3\geq 8,$ it implies

$$\mathbb{E}[(\hat{\mu}_{k,n} - \mu_{k})^{2}\mathbb{I}\left\{\xi\right\}]$$

$$\leq \frac{1}{(n-2K)^{2}} \left(n\Sigma + \frac{3B}{\lambda_{k}}\sqrt{n} + \frac{3C + 2K\Sigma}{\lambda_{k}}n^{1/4} + \frac{K(B+C)^{3}}{\lambda_{k}}\left(\frac{1}{2\Sigma} + 1 + \frac{1}{8\Sigma} + \frac{1}{2\Sigma^{2}} + \frac{1}{\Sigma}\right) + \frac{4\delta n^{2}K}{\lambda_{k}}\left(\Sigma + 2(B+C) + \frac{(B+C)^{2}}{\Sigma}\right)\right)$$

$$\leq \frac{1}{(n-2K)^{2}} \left(n\Sigma + \frac{3B}{\lambda_{k}}\sqrt{n} + \frac{3C + 2K\Sigma}{\lambda_{k}}n^{1/4} + \frac{K(B+C)^{3}}{\lambda_{k}}\left(\frac{1}{2\Sigma^{2}} + \frac{13}{8}\Sigma + 1\right) + \frac{4\delta n^{2}K}{\lambda_{k}}\left(\Sigma + 2(B+C) + \frac{(B+C)^{2}}{\Sigma}\right)\right)$$

$$\leq \frac{1}{(n-2K)^{2}} \left(n\Sigma + \frac{3B}{\lambda_{k}}\sqrt{n} + \frac{3C + 2K\Sigma}{\lambda_{k}}n^{1/4} + K\frac{(B+C)^{3}}{\lambda_{k}}\left(\frac{1}{\Sigma^{2}} + 8\right) + \frac{4\delta n^{2}K}{\lambda_{k}}\left(\Sigma + 2(B+C) + \frac{(B+C)^{2}}{\Sigma}\right)\right).$$

Now note that, as $\delta = n^{-7/2}$ and $n \ge 4K$

$$\mathbb{E}[(\hat{\mu}_{k,n} - \mu_{k})^{2}\mathbb{I}\left\{\xi\right\}] \leq \frac{1}{(n-2K)^{2}} \left(n\Sigma + \frac{3B}{\lambda_{k}}\sqrt{n} + \frac{3C + 2K\Sigma}{\lambda_{k}}n^{1/4} + K\frac{(B+C)^{3}}{\lambda_{k}}(\frac{1}{\Sigma^{2}} + 8) + \frac{4K\Sigma}{n^{3/2}\lambda_{k}}(1 + \frac{B+C}{\Sigma})^{2}\right)$$

$$\leq \left(\frac{1}{n^{2}} + \frac{8K}{n^{3}}\right) \left(n\Sigma + \frac{3B}{\lambda_{k}}\sqrt{n} + \frac{3C + 2K\Sigma}{\lambda_{k}}n^{1/4} + K\frac{(B+C)^{3}}{\lambda_{k}}(\frac{1}{\Sigma^{2}} + 8) + \frac{8K\Sigma}{n^{3/2}\lambda_{k}}(B+C)^{2}(1 + \frac{1}{\Sigma^{2}})\right)$$

$$\leq \frac{\Sigma}{n} + \frac{8K\Sigma}{n^{2}} + \frac{3}{n^{2}} \left(\frac{3B}{\lambda_{k}}\sqrt{n} + \frac{3C + 2K\Sigma}{\lambda_{k}}n^{1/4} + K\frac{(B+C)^{3}}{\lambda_{k}}(\frac{1}{\Sigma^{2}} + 8) + \frac{8K\Sigma}{n^{3/2}\lambda_{k}}(B+C)^{2}(1 + \frac{1}{\Sigma^{2}})\right)$$

$$\leq \frac{\Sigma}{n} + \frac{9B}{n^{3/2}\lambda_{k}} + \frac{8K\Sigma}{n^{2}} + \frac{3}{n^{7/4}\lambda_{k}} \left(3C + 2K\Sigma + K(B+C)^{3}(1 + \Sigma)(\frac{1}{\Sigma^{2}} + 8)\right)$$

$$\leq \frac{\Sigma}{n} + \frac{9B}{n^{3/2}\lambda_{k}} + \frac{8K\Sigma}{n^{2}} + \frac{3}{n^{7/4}\lambda_{k}} \left(K(B+C)^{3}(1 + \Sigma)(\frac{1}{\Sigma^{2}} + 13)\right)$$

$$\leq \frac{\Sigma}{n} + \frac{9B}{n^{3/2}\lambda_{min}} + 3K(B+C)^{3}(1 + \Sigma)(\frac{1}{\Sigma^{2}} + 21)\frac{1}{n^{7/4}\lambda_{min}}$$

again since $B + C \ge 1$.

Finally, combining that with Lemma 6 gives us for the regret:

$$R_n(\mathcal{A}_B) \leq \frac{9B}{n^{3/2}\lambda_{\min}} + 3K\frac{(B+C)^3}{n^{7/4}\lambda_{\min}} (\frac{1}{\Sigma^2} + 21)(1+\Sigma) + 2c_1n^2K\delta(1+\log(c_2/2nK\delta)).$$

By taking $\delta = n^{-7/2}$ and recalling the bounds on B and C in Appendix B.3.2, we obtain:

$$R_n(\mathcal{A}_B) \le \frac{9B}{n^{3/2}\lambda_{\min}} + 3K \frac{(B+C)^3}{n^{7/4}\lambda_{\min}} (\frac{1}{\Sigma^2} + 21)(1+\Sigma) + 7c_1(c_2+1)K \log(n)n^{-3/2}$$

$$\le \frac{76400c_1(c_2+1)K^2 \log(n)^2}{\lambda_{\min}n^{3/2}} + O\left(\frac{\log(n)^6K^7}{n^{7/4}\lambda_{\min}}\right).$$

Appendix C. Regret Bound for Gaussian Distributions

Here we report the proof of Lemma 3 which implies that when the distributions of the arms are Gaussian, bounding the regret of the B-AS algorithm does not require upper-bounding the number of pulls $T_{k,n}$ (it can be bounded only by using a lower bound on the number of pulls).

Let $\{X_t\}_{t\geq 1}$ be a sequence of i.i.d. random variables drawn from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Write $\hat{m}_t = \frac{1}{t} \sum_{i=1}^t X_i$ and $\hat{s}_t^2 = \frac{1}{t-1} \sum_{i=1}^t (X_i - \hat{m}_t)^2$ for the empirical mean and variance of the first t samples. Before proving Lemma 3, we recall a property of the normal distribution (see e.g., [4]).

Proposition 3. Let X_1, \ldots, X_t be t i.i.d. Gaussian random variables. Then their empirical mean $\hat{m}_t = \frac{1}{t} \sum_{i=1}^t X_i$ and empirical variance $\hat{s}_t^2 = \frac{1}{t-1} \sum_{i=1}^t (X_i - \hat{m}_t)^2$ are independent of each other.

Based only on the well-known t=2 case (i.e., that X_1+X_2 and $|X_1-X_2|$ are independent), we can derive a somewhat stronger result that is used in the proof of Lemma 3, showing that for Gaussian distributions, the empirical mean \hat{m}_t built on t i.i.d. samples is independent from the sequence of standard deviations $(\hat{s}_2, ..., \hat{s}_t)$ (not only from \hat{s}_t^2).

We first derive a general result showing that for Gaussian distributions, the empirical mean \hat{m}_t built on t i.i.d. samples is independent from the sequence of standard deviations $\hat{s}_2, \ldots, \hat{s}_t$.

Lemma 7. Let \mathcal{F}_t be the σ -algebra generated by the sequence of random variables $\hat{s}_2, \ldots, \hat{s}_t$. Then for all $t \geq 2$,

$$\hat{m}_t | \mathcal{F}_t \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{t}\right).$$

To prove Lemma 7, we need the following technical lemma:

Lemma 8. We have

$$\hat{s}_{t+1}^2 = \frac{t-1}{t} \hat{s}_t^2 + \frac{1}{t+1} (X_{t+1} - \hat{m}_t)^2.$$

Note that this statement is deterministic, it holds for any process or sequence.

Proof. We have for $t \geq 2$

$$\hat{s}_{t+1}^{2} = \frac{1}{t} \sum_{i=1}^{t+1} (X_{i} - \hat{m}_{t+1})^{2}$$

$$= \frac{1}{t} \sum_{i=1}^{t} (X_{i} - \hat{m}_{t+1} + \hat{m}_{t} - \hat{m}_{t})^{2} + \frac{1}{t} (X_{t+1} - \hat{m}_{t+1})^{2}$$

$$= \frac{1}{t} \sum_{i=1}^{t} (X_{i} - \hat{m}_{t})^{2} + \frac{1}{t} (X_{t+1} - \hat{m}_{t+1})^{2} + (\hat{m}_{t} - \hat{m}_{t+1})^{2}$$

$$= \frac{1}{t} \sum_{i=1}^{t} (X_{i} - \hat{m}_{t})^{2} + \frac{t}{(t+1)^{2}} (X_{t+1} - \hat{m}_{t})^{2} + \frac{1}{(t+1)^{2}} (X_{t+1} - \hat{m}_{t})^{2}$$

$$= \frac{1}{t} \sum_{i=1}^{t} (X_{i} - \hat{m}_{t})^{2} + \frac{1}{t+1} (X_{t+1} - \hat{m}_{t})^{2},$$

which finishes the proof.

From Lemma 8 we deduce by induction that for any $t \ge 2$ there exists a sequence of non-negative real numbers $\{a_{1,t}, a_{2,t}, \dots, a_{t,t}\}$ such that

$$\hat{s}_t^2 = a_{1,t}\hat{s}_2^2 + \sum_{i=2}^{t-1} a_{i,t}(X_{i+1} - \hat{m}_i)^2.$$

Proof. We prove the statement by induction.

The base of the induction (t=2) is directly implied by the specific properties of Gaussian distributions (Proposition 3). In fact, \hat{m}_2 is distributed as $\mathcal{N}(\mu, \sigma^2/2)$ and \hat{m}_2 and \hat{s}_2 are independent.

Now we focus on the inductive step. For any $t \geq 2$, let \mathcal{G}_t be the σ -algebra generated by the random variables \hat{s}_2^2 and $\{(X_{i+1} - \hat{m}_i)^2\}_{2 \leq i \leq t-1}$. The recursive definition of the empirical variance in Lemma 8 immediately implies that the knowledge of $\{\hat{s}_2, \dots, \hat{s}_t\}$ is equivalent to the knowledge of \hat{s}_2^2 and $\{(X_{i+1} - \hat{m}_i)^2\}_{2 \leq i \leq t-1}$ and thus $\mathcal{F}_t = \mathcal{G}_t$. We assume (inductive hypothesis)

$$\hat{m}_t | \mathcal{G}_t \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{t}\right),$$
 (C.1)

and we now show that (C.1) also holds for t+1. Let $U=X_{t+1}-\hat{m}_t$ and $V=\hat{m}_{t+1}-\mu$. Note that V can be written as $V=\frac{t}{t+1}(\hat{m}_t-\mu)+\frac{1}{t+1}(X_{t+1}-\mu)$. Since samples are i.i.d., X_{t+1} is independent from (X_1,\ldots,X_t) and

$$X_{t+1} | \mathcal{G}_t \sim \mathcal{N}(\mu, \sigma^2)$$

and thus X_{t+1} is also conditionally independent of \hat{m}_t given \mathcal{G}_t . This implies that X_{t+1} and \hat{m}_t are jointly Gaussian given \mathcal{G}_t (two random variables that are Gaussian and independent are jointly Gaussian, see [9] or also http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Joint_normality). This fact combined with the definition of U and V implies that U and V are conditionally jointly-Gaussian variables with zero conditional mean given \mathcal{G}_t (they are jointly-Gaussian because they can be written as two independent linear combinations of the random variables $X_{t+1} - \mu$ and $\hat{m}_t - \mu$ given \mathcal{G}_t , see [9] or also http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Affine_transformation). Furthermore, we can show that they are also conditionally uncorrelated given \mathcal{G}_t since

$$\mathbb{E}\Big[UV|\mathcal{G}_{t}\Big] = \mathbb{E}\Big[\Big(X_{t+1} - \hat{m}_{t}\Big)\Big(\frac{1}{t+1}X_{t+1} + \frac{t}{t+1}\hat{m}_{t} - \mu\Big)\Big|\mathcal{G}_{t}\Big]$$

$$= \mathbb{E}\Big[\Big((X_{t+1} - \mu) - (\hat{m}_{t} - \mu)\Big)\Big(\frac{1}{t+1}(X_{t+1} - \mu) + \frac{t}{t+1}(\hat{m}_{t} - \mu)\Big)\Big|\mathcal{G}_{t}\Big]$$

$$= \frac{1}{t+1}\sigma^{2} - \frac{t}{t+1}\frac{\sigma^{2}}{t} = 0.$$

As a result, U and V are conditionally independent given \mathcal{G}_t and

$$(\hat{m}_{t+1} - \mu) | \mathcal{G}_{t+1} = (\hat{m}_{t+1} - \mu) | \{\mathcal{G}_t, (X_{t+1} - \hat{m}_t)^2\} = (\hat{m}_{t+1} - \mu) | \{\mathcal{G}_t, U^2\} = V | \{\mathcal{G}_t, U^2\} = V | \mathcal{G}_t.$$

Since the induction assumption is verified, we know that $\mathbb{E}[V|\mathcal{G}_t] = 0$ and $\mathbb{V}[V|\mathcal{G}_t] = (\frac{t}{t+1})^2 \frac{\sigma^2}{t} + (\frac{1}{t+1})^2 \sigma^2 = \frac{\sigma^2}{t+1}$. Finally, we deduce that

$$\hat{m}_{t+1} | \mathcal{G}_{t+1} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{t+1}\right),$$

which concludes the proof since $\mathcal{G}_{t+1} = \mathcal{F}_{t+1}$.

We now study an adaptive algorithm that computes the empirical average \hat{m}_t and that at each time t decides whether to stop collecting samples or not on the basis of the sequence of empirical standard deviations $\hat{s}_2, \ldots, \hat{s}_t$ observed so far. Let $T \geq 2$ be a integer-valued random variable, which is a stopping time with respect to \mathcal{F}_t . This means that the decision of whether to stop at any time before t+1 (the event $\{T \leq t\}$) only depends on the previous empirical standard deviations $\hat{s}_2, \ldots, \hat{s}_t$. From an immediate application of Lemma 7 we obtain

$$\begin{split} \mathbb{E}[(\hat{m}_T - \mu)^2] &= \sum_{t \geq 2} \mathbb{E}[(\hat{m}_t - \mu)^2 | T = t] \mathbb{P}[T = t] \\ &= \sum_{t \geq 2} \mathbb{E}[\mathbb{E}[(\hat{m}_t - \mu)^2 | \mathcal{F}_t, T = t] | T = t] \mathbb{P}[T = t] \\ &= \sum_{t \geq 2} \mathbb{E}[\mathbb{E}[(\hat{m}_t - \mu)^2 | \mathcal{F}_t] | T = t] \mathbb{P}[T = t] = \sum_{t \geq 2} \frac{\sigma^2}{t} \mathbb{P}[T = t] = \sigma^2 \mathbb{E}\left[\frac{1}{T}\right]. \end{split}$$

The previous result seamlessly extends to the general multi-armed bandit allocation strategies considered in Section 3 and 4.

Proof of Lemma 3. Let us now consider algorithms CH-AS and B-AS. For any arm k, the event $\{T_{k,n} > t\}$ depends on the σ -algebra $\mathcal{F}_{k,t}$ (generated by the sequence of empirical variances of the first t samples of arm k) and also on the "environment" \mathcal{E}_{-k} (generated by all the samples of other arms). Since the samples of arm k are independent from \mathcal{E}_{-k} , we deduce that by conditioning on \mathcal{E}_{-k} Lemma 7 still applies and

$$\mathbb{E}[(\hat{\mu}_{k,n} - \mu)^2] = \mathbb{E}_{\mathcal{E}_{-k}} \left[\mathbb{E}[(\hat{\mu}_{k,n} - \mu)^2 | \mathcal{E}_{-k}] \right] = \sigma_k^2 \mathbb{E}_{\mathcal{E}_{-k}} \left[\mathbb{E} \left[\frac{1}{T_{k,n}} | \mathcal{E}_{-k} \right] \right] = \sigma_k^2 \mathbb{E} \left[\frac{1}{T_{k,n}} | \mathcal{E}_{-k} \right]$$

We now report the proof of Theorem 3.

Proof of Theorem 3. We recall Lemma 3 and decompose the loss using the definition of $\xi = \xi_{K,n}^B(\delta)$ in order to obtain

$$L_{k,n} = \sigma_k^2 \mathbb{E} \Big[\frac{1}{T_{k,n}} \Big] = \sigma_k^2 \mathbb{E} \Big[\frac{1}{T_{k,n}} \mathbb{I} \left\{ \xi \right\} \Big] + \sigma_k^2 \mathbb{E} \Big[\frac{1}{T_{k,n}} \mathbb{I} \left\{ \xi^c \right\} \Big].$$

From the bound in Equation (B.20), we have (since $n \geq 5K$)

$$\begin{split} \sigma_k^2 \mathbb{E} \Big[\frac{1}{T_{k,n}} \mathbb{I} \left\{ \xi \right\} \Big] &\leq \max_{\xi} \left[\frac{\sigma_k^2}{T_{k,n}} \right] \\ &\leq \frac{\Sigma}{n - 2K} + \frac{B}{n^{1/2}(n - 2K)} + \frac{C}{n^{3/4}(n - 2K)} \\ &\leq \frac{\Sigma}{n} + \frac{4K\Sigma}{n^2} + \frac{2B}{n^{3/2}} + \frac{2C}{n^{7/4}} \\ &\leq \frac{\Sigma}{n} + \frac{4K\Sigma}{n^2} + \frac{12 \times 10^3}{n^{3/2}} K^2 c_1 (c_2 + 1) (\log n)^2 + \frac{14 \times 10^3}{n^{7/4}} K^2 c_1 (c_2 + 1) (\log n)^2 \\ &\leq \frac{\Sigma}{n} + \frac{12.001 \times 10^3}{n^{3/2}} K^2 c_1 (c_2 + 1) (\log n)^2 + \frac{14 \times 10^3}{n^{7/4}} K^2 c_1 (c_2 + 1) (\log n)^2 \\ &\leq \frac{\Sigma}{n} + \frac{26.001 \times 10^3}{n^{3/2}} K^2 c_1 (c_2 + 1) (\log n)^2. \end{split}$$

$$(C.2)$$

where we use the bounds on B and C in Appendix B.3.2. Using the fact that $\delta = n^{-7/2}$ and $T_{k,n} \geq 2$, and by Lemma 4 that tells us $\mathbb{P}[\xi^c] \leq 2nK\delta$, we may write

$$\sigma_k^2 \mathbb{E}\left[\frac{1}{T_{k_n}} \mathbb{I}\left\{\xi^c\right\}\right] \le K \sigma_k^2 n^{-5/2} \le c_1 c_2 K n^{-5/2}. \tag{C.3}$$

Finally, combining Equations C.2 and C.3, and recalling the definition of regret, we have

$$R_{n}(\mathcal{A}_{B}) \leq \frac{26.001 \times 10^{3}}{n^{3/2}} K^{2} c_{1} (c_{2} + 1) (\log n)^{2} + c_{1} c_{2} K n^{-5/2}$$

$$\leq \frac{26.002 \times 10^{3}}{n^{3/2}} K^{2} c_{1} (c_{2} + 1) (\log n)^{2}$$

$$\leq \frac{105 \times 10^{3} \bar{\Sigma}}{n^{3/2}} K^{2} (\log n)^{2},$$
(C.5)

since $c_1 = 2\bar{\Sigma}$ and $c_2 = 1$.