View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by INRIA a CCSD electronic archive server

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

Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, Peter Auer

To cite this version:

Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, Peter Auer. Upper-Confidence-Bound Algorithms for Active Learning in Multi-Armed Bandits. ALT - the 22nd conference on Algorithmic Learning Theory, Oct 2011, Espoo, Finland. hal-00659696

HAL Id: hal-00659696

https://hal.inria.fr/hal-00659696

Submitted on 13 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Alexandra Carpentier[1], Alessandro Lazaric[1], Mohammad Ghavamzadeh[1], Rémi Munos[1], and Peter Auer[2]

[1] SequeL team, INRIA Lille - Nord Europe, Team SequeL, France
[2] University of Leoben, Franz-Josef-Strasse 18, 8700 Leoben, Austria

Abstract. In this paper, we study the problem of estimating the mean values of all the arms uniformly well in the multi-armed bandit setting. If the variances of the arms were known, one could design an optimal sampling strategy by pulling the arms proportionally to their variances. However, since the distributions are not known in advance, we need to design adaptive sampling strategies to select an arm at each round based on the previous observed samples. We describe two strategies based on pulling the arms proportionally to an upper-bound on their variances and derive regret bounds for these strategies. We show that the performance of these allocation strategies depends not only on the variances of the arms but also on the full shape of their distributions.

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 (Cohn et al., 1996, Castro et al., 2005), sampling and Monte-Carlo methods (Étoré and Jourdain, 2010), and optimal experimental design (Fedorov, 1972, Chaudhuri and Mykland, 1995). A particular instance of this problem is introduced in Antos et al. (2010) 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 Eq. 4) that is the difference between the expected quadratic estimation error of the algorithm and the error of the optimal allocation.

Antos et al. (2010) 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). 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.[3] 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.

The main objective of this paper is to investigate this issue and identify under which conditions it 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 $\lambda_{\min}$, 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 an empirical Bernstein's inequality, and has a better performance (in terms of the number of pulls) in targeting the optimal allocation strategy without any dependency on $\lambda_{\min}$. However, moving from the number of pulls to the regret causes the inverse dependency on $\lambda_{\min}$ to appear in the bound again. We show that this might be due to the specific shape of the distributions $\{\nu_k\}_{k=1}^K$ and derive a regret bound independent of $\lambda_{\min}$ for the case of Gaussian arms.

[3] The 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.

– We show empirically that while the performance of CH-AS depends on $\lambda_{\min}$ 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 $\lambda_{\min}$ 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 shape of the distributions, and not only their variance, impacts 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 $k = 1, \dots, K$ is characterized by a distribution ν_k with mean μ_k and variance σ_k^2. At each round $t \geq 1$, 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 arm k, the algorithm $\mathcal{A}$ returns the empirical estimates $\hat{\mu}_{k,n} = \dfrac{1}{T_{k,n}} \displaystyle\sum_{t=1}^{T_{k,n}} X_{k,t}$, where $X_{k,t}$ denotes the sample received when pulling arm k for the t-th time. The accuracy of the estimation at each arm k is measured according to its expected squared estimation error, or loss

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

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

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

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 this case, if an arm k is pulled a fixed number of times $T^*_{k,n}$, its loss is computed as[4]

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

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)

[4] 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.

$T^*_{k,n} = \frac{\sigma_k^2 n}{\sum_{i=1}^K \sigma_i^2}$ 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 the loss incurred by the learner and the optimal loss $L_n(\mathcal{A}^*)$:

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

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., Audibert et al. 2010, Bubeck et al. 2011). 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.

3 Allocation Strategy Based on Chernoff-Hoeffding UCB

3.1 The CH-AS Algorithm

The first algorithm introduced in this paper, 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. The pseudo-code of the CH-AS algorithm $\mathcal{A}_{CH}$ is given in Fig. 1. It takes a confidence parameter δ as input and after n pulls returns an empirical mean $\hat{\mu}_{q,n}$ for each arm q. At each time step t, the algorithm computes the empirical mean $\hat{\mu}_{q,t}$ and variance $\hat{\sigma}^2_{q,t}$ of each arm q as[5]

$$\hat{\mu}_{q,t} = \frac{1}{T_{q,t}} \sum_{i=1}^{T_{q,t}} X_{q,i} \qquad \text{and} \qquad \hat{\sigma}^2_{q,t} = \frac{1}{T_{q,t}} \sum_{i=1}^{T_{q,t}} X^2_{q,i} - \hat{\mu}^2_{q,t}\ , \tag{5}$$

where $X_{q,i}$ is the i-th sample of ν_k and $T_{q,t}$ is the number of pulls allocated to arm q up to time t. After pulling each arm twice (rounds $t = 1$ to $2K$),

[5] Notice that this is a biased estimator of the variance.

Input: parameter δ
Initialize: Pull each arm twice
for $t = 2K+1, \ldots, n$ **do**
  Compute $B_{q,t} = \frac{1}{T_{q,t}}\Big(\hat{\sigma}^2_{q,t} + 5\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}}\Big)$ 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: $\hat{\mu}_{q,n}$ for all arms $1 \le q \le K$

Fig. 1. The pseudo-code of the CH-AS algorithm. The empirical variances $\hat{\sigma}^2_{q,t}$ are computed from Eq. 5.

from round $t = 2K+1$ on, the algorithm computes the $B_{q,t}$ values based on a Chernoff-Hoeffding's bound on the variances of the arms:

$$B_{q,t} = \frac{1}{T_{q,t}}\Big(\hat{\sigma}^2_{q,t} + 5\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}}\Big),$$

and then pulls the arm with the largest $B_{q,t}$.

3.2 Regret Bound and Discussion

Before reporting a regret bound for CH-AS, 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 support of the distributions $\{\nu_k\}_{k=1}^K$ are in $[0,1]$. For any $\delta > 0$, any arm $1 \le k \le K$, and any time $1 \le t \le n$, with probability $1 - 4nK\delta$, the number of pulls by the CH-AS algorithm satisfies*

$$-\frac{2}{\Sigma^2\lambda_{\min}^{3/2}}\sqrt{n\log(1/\delta)} \le T_{k,n} - T^*_{k,n} \le \frac{2(K-1)}{\Sigma^2\lambda_{\min}^{3/2}}\sqrt{n\log(1/\delta)}\,. \tag{6}$$

Proof. Let ξ be the event

$$\xi = \bigcap_{1\le k\le K,\ 1\le t\le n}\left\{\Big|\Big(\frac{1}{t}\sum_{i=1}^t X^2_{k,i} - \big(\frac{1}{t}\sum_{i=1}^t X_{k,i}\big)^2\Big) - \sigma^2_k\Big| \le 5\sqrt{\frac{\log(1/\delta)}{2t}}\right\}. \tag{7}$$

It can be shown using Hoeffding's inequality that $\Pr(\xi) \ge 1 - 4nK\delta$. Several of the following statements will be proved on this event. We divide the proof of this lemma into the following three steps.

Step 1. Mechanism of the algorithm. From Chernoff-Hoeffding's inequality (applied to $X_{k,t}$ and $X^2_{k,t}$), one may prove that for any $\delta > 0$, there exists an event ξ with probability at least $1 - nK\delta$ (a more detailed definition of this event and its probability is available at the longer version of the paper (Carpentier et al., 2011)), such that on ξ, for all $t \le n$ and $q \le K$, we have

$$|\hat{\sigma}^2_{q,t} - \sigma^2_q| \leq 5\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}}\ ,$$

and the following upper and lower bounds for $B_{q,t}$

$$\frac{\sigma^2_q}{T_{q,t}} \leq B_{q,t} \leq \frac{1}{T_{q,t}}\left(\sigma^2_q + 10\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}}\right). \tag{8}$$

Let t be the time when arm k is pulled for the last time, i.e., $T_{k,t} = T_{k,n} - 1$ and $T_{k,(t+1)} = T_{k,n}$. Since $\mathcal{A}_{CH}$ chooses to pull arm k at time t, for any arm $p \neq k$, we have $B_{p,t} \leq B_{k,t}$. From Eq. 8 and the fact that $T_{k,t} = T_{k,n} - 1$, we obtain

$$B_{k,t} \leq \frac{1}{T_{k,t}}\left(\sigma^2_k + 10\sqrt{\frac{\log(1/\delta)}{2T_{k,t}}}\right) = \frac{1}{T_{k,n}-1}\left(\sigma^2_k + 10\sqrt{\frac{\log(1/\delta)}{2(T_{k,n}-1)}}\right). \tag{9}$$

Using the lower-bound in Eq. 8 and the fact that $T_{p,t} \leq T_{p,n}$, we derive a lower-bound for $B_{p,t}$ as

$$B_{p,t} \geq \frac{\sigma^2_p}{T_{p,t}} \geq \frac{\sigma^2_p}{T_{p,n}}\ . \tag{10}$$

Combining the condition $B_{p,t} \leq B_{k,t}$ with Eqs. 9 and 10, we obtain

$$\frac{\sigma^2_p}{T_{p,n}} \leq \frac{1}{T_{k,n}-1}\left(\sigma^2_k + 10\sqrt{\frac{\log(1/\delta)}{2(T_{k,n}-1)}}\right)\ . \tag{11}$$

Note that at this point there is no dependency on t, and thus, Eq. 11 holds with probability $1 - 4nK\delta$ (this is because Eq. 11 is defined on the event ξ) for any pair of arms p and k.

Step 2. Lower bound on $T_{p,n}$. If an arm p is under-pulled, i.e., $T_{p,n} < T^*_{p,n}$, then from the constraint $\sum_k T_{k,n} = n$ and the definition of the optimal allocation, we may deduce that there must be at least one arm k that is over-pulled, i.e., $T_{k,n} > T^*_{k,n}$. Using the definition of the optimal allocation $T^*_{k,n} = n\lambda_k = n\sigma^2_k/\Sigma$ and the fact that $T_{k,n} \geq T^*_{k,n} + 1$, Eq. 11 may be rewritten as

$$\frac{\sigma^2_p}{T_{p,n}} \leq \frac{1}{T^*_{k,n}}\left(\sigma^2_k + 10\sqrt{\frac{\log(1/\delta)}{2T^*_{k,n}}}\right) = \frac{\Sigma}{n} + \frac{5\sqrt{2\log(1/\delta)}}{(n\lambda_k)^{3/2}} \leq \frac{\Sigma}{n} + \frac{5\sqrt{2\log(1/\delta)}}{(n\lambda_{\min})^{3/2}}. \tag{12}$$

By reordering the terms in Eq. 12, we obtain a lower bound on $T_{p,n}$ as

$$T_{p,n} \geq \frac{\sigma^2_p}{\frac{\Sigma}{n} + \frac{5\sqrt{2\log(1/\delta)}}{(n\lambda_{\min})^{3/2}}} \geq T^*_{p,n} - \frac{5\sqrt{2n\log(1/\delta)}}{4\Sigma^2\lambda^{3/2}_{\min}} \geq T^*_{p,n} - \frac{2}{\Sigma^2\lambda^{3/2}_{\min}}\sqrt{n\log(1/\delta)}, \tag{13}$$

where in the second inequality we used $1/(1+x) \geq 1 - x$ (for $x > -1$) and $\sigma^2_p \leq 1/4$, and in the last passage we used $5\sqrt{2}/4 < 2$. Note that the lower bound holds w.h.p. for any arm p.

Step 3. Upper bound on $T_{p,n}$. Using Eq. 33 and the fact that $\sum_k T_{k,n} = n$, we obtain

$$T_{p,n} = n - \sum_{k \neq p} T_{k,n} \leq T^*_{p,n} + \frac{2(K-1)}{\Sigma^2 \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} \,. \tag{14}$$

The claim follows by combining the upper and lower bounds. □

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 support of the distributions $\nu_1, \ldots, \nu_K$ are in $[0,1]$. For any $n \geq 0$, the regret of $\mathcal{A}_{CH}$, when it runs with the parameter $\delta = n^{-5/2}$, is bounded as*

$$R_n(\mathcal{A}_{CH}) \leq \frac{28K\sqrt{\log n}}{n^{3/2}\ \Sigma\ \lambda_{\min}^{5/2}} + O\Big(\frac{\log n}{n^2}\Big) \,. \tag{15}$$

For space limitations, we only report a sketch of the proof here, the full proof is provided in the longer version of the paper (Carpentier et al., 2011).

Proof (Sketch). Eq. 3 indicates that the more an arm is pulled, the more its estimation error becomes small. However, this is not true in general because $T_{k,n}$ is a random variable that depends on the actual received rewards, and thus, $L_{k,n} = \mathbb{E}_{\nu_k}\big[(\mu_k - \hat{\mu}_{k,n})^2\big]$ does not satisfy Eq. 3. Nevertheless, we have the property that for any arm k, the number of pulls $T_{k,n}$ is a stopping time w.r.t. the filtration induced by the rewards received for arm k. Hence, by applying the result of Lemma 2 in Antos et al. (2010) (a form of the Wald's theorem), one may derive [6]

$$L_{k,n}(\xi) = \mathbb{E}\big[(\mu_k - \hat{\mu}_{k,n})^2 \mathbb{I}\{\xi\}\big] \leq \frac{1}{\underline{T}_{k,n}^2} \mathbb{E}\big[\sum_{t=1}^{T_{k,n}} (\mu_k - X_{k,t})^2\big] = \frac{\sigma_k^2 \mathbb{E}(T_{k,n})}{\underline{T}_{k,n}^2} \,, \tag{16}$$

where $\underline{T}_{k,n}$ is a lower-bound for $T_{k,n}$ on ξ. From this bound, one can use Lemma 4, which provides both upper and lower-bounds for $T_{k,n}$ on the event ξ to deduce that $L_{k,n}(\xi) = \frac{\sigma_k^2}{T^*_{k,n}} + O\big(n^{-3/2}\log(1/\delta)\big)$ and $L_{k,n}(\xi^c) \leq 1$ (which is obvious). The claim follows by setting $\delta = n^{-5/2}$. □

Remark 1. As discussed in Sec. 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, Thm. 1 shows that the CH-AS algorithm provides a uniformly accurate estimation of the expected values of the arms with a regret R_n of order $\widetilde{O}(n^{-3/2})$. This regret rate is the same as the one for the GAFS-MAX algorithm (Antos et al., 2010).

[6] The total loss $L_{k,n}$ is decomposed as $L_{k,n} = L_{k,n}(\xi) + L_{k,n}(\xi^c)$.

Remark 2. In addition to a linear dependency on the number of arms K, the bound also displays an inverse dependency on the smallest proportion $\lambda_{\min}$. As a result, the bound scales poorly when an arm has a very small variance relative to the other arms (i.e., $\sigma_k \ll \Sigma$). Note that GAFS-MAX has also a similar dependency on the inverse of $\lambda_{\min}$, although a precise comparison is not possible due to the fact that Antos et al. (2010) do not explicitly report the multiplicative constants in their regret bound. Moreover, Thm. 1 holds for any n whereas the regret bound in Antos et al. (2010) requires a condition $n \geq n_0$, where n_0 is a constant that scales with $\lambda_{\min}^{-1}$. 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 4 that the inverse dependency on $\lambda_{\min}$ appears in the bound on the number of pulls and then is propagated to the regret bound. We now show with a simple example that this dependency is not an artifact of the analysis and is intrinsic in the performance of the algorithm. Consider a two-arm problem with $\sigma_1^2 = 1$ and $\sigma_2^2 = 0$. Here 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 $\lambda_{\min} = 0$, which makes the bound in Thm. 1 vacuous. This does not mean that CH-AS has a linear regret, it indicates that it minimizes the regret with a poorer rate (see Sec. A.3 in Carpentier et al. 2011, for a sketch of the proof). In fact, the Chernoff-Hoeffding's bound used in the upper-confidence term forces the algorithm to pull the arm with zero variance at least $\widetilde{O}(n^{2/3})$ times, which results in under-pulling the first arm by the same amount, and thus, in worsening its estimation. It can be shown that the resulting regret has the rate $\widetilde{O}(n^{-4/3})$ and no dependency on $\lambda_{\min}$. So, it still decreases to zero, but with a slower rate than the one in Thm. 1. Merging these two results, we deduce that the regret of CH-AS is in fact $R_n \leq \min\left\{\tilde{O}(n^{-3/2}\lambda_{\min}^{-5/2}), \tilde{O}(n^{-4/3})\right\}$. We will further study the behavior of CH-AS, i.e., how its regret changes with n, in Sec. 5.1.

The reason for the poor performance in Lemma 4 is that Chernoff-Hoeffding's inequality is not tight for small-variance random variables. In Sec. 4, we propose an algorithm based on an empirical Bernstein's inequality, which is tighter for small-variance random variables, and prove that this algorithm under-pulls all the arms by *at most* $\widetilde{O}(n^{1/2})$, without a dependency on $\lambda_{\min}$ (see Eqs. 19 and 20).

4 Allocation Strategy Based on Bernstein UCB

In this section, we present another UCB-like algorithm, called *Bernstein Allocation Strategy* (B-AS), based on a Bernstein's inequality for the variances of the arms, that enables us to improve the bound on $|T_{k,n} - T^*_{k,n}|$ by removing the inverse dependency on $\lambda_{\min}$ (compare the bounds in Eqs. 19 and 20 to the one for CH-AS in Eq. 26). 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 $\lambda_{\min}$. We further investigate this issue by showing that the way the bound of

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

Fig. 2. The pseudo-code of the B-AS algorithm. The empirical variances $\hat{\sigma}_{k,t}$ are computed according to Eq. 17

the number of pulls translates into a regret bound depends on the specific distributions of the arms. In fact, when the reward distributions are Gaussian, we can exploit the property that the empirical variance $\hat{\sigma}_{k,t}$ is independent of the empirical mean $\hat{\mu}_{k,t}$, and show that the regret of B-AS no longer depends on $\lambda_{\min}^{-1}$. The numerical simulations in Sec. 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 (empirical Bernstein's inequality), reported in Maurer and Pontil (2009) (a similar bound can be found in Audibert et al. 2009), on the variance of each arm. Like in the previous section, the arm sampling strategy is proportional to those bounds. The B-AS algorithm, $\mathcal{A}_B$, is described in Fig. 2. It requires three parameters as input (see Remark 4 in Sec. 4.2 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 properly tuning these parameters. The algorithm is similar to CH-AS except that the bounds $B_{q,t}$ on each arm are computed as

$$B_{q,t} = \frac{1}{T_{q,t}}\left(\hat{\sigma}^2_{q,t} + 2b\hat{\sigma}_{q,t}\sqrt{\frac{1}{T_{q,t}}} + b^2\frac{1}{T_{q,t}}\right),$$

where $b = 2\sqrt{2c_1 \log(c_2/\delta)\log(2/\delta)} + \frac{\sqrt{2c_1\delta(1+c_2+\log(c_2/\delta))}}{(1-\delta)} n^{1/2}$ and[7]

$$\hat{\sigma}^2_{k,t} = \frac{1}{T_{k,t}-1}\sum_{i=1}^{T_{k,t}}(X_{k,i} - \hat{\mu}_{k,t})^2, \qquad \text{with} \qquad \hat{\mu}_{k,t} = \frac{1}{T_{k,t}}\sum_{i=1}^{T_{k,t}} X_{k,i}\,. \tag{17}$$

[7] We consider the unbiased estimator of the variance here.

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 small variances (Berstein's bound is tighter than Chernoff-Hoeffding's bound for distributions with small variance). 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 $1 \leq k \leq K$ and any $\epsilon > 0$,*

$$\mathbb{P}_{X\sim\nu_k}(|X-\mu_k| \geq \epsilon) \leq c_1 \exp(-c_2\epsilon^2) \ . \tag{18}$$

We first state a bound in Lemma 2 on the difference between the B-AS and optimal allocation strategies.

Lemma 2. *Under Assumption 1 and for any $\delta > 0$, when the B-AS algorithm runs with parameters c_1, c_2, and δ, with probability at least $1 - 2nK\delta$, we have*

$$T_{p,n} \geq T^*_{p,n} - K\lambda_p \left[1 + \frac{16a\sqrt{\log(2/\delta)}}{\Sigma}\left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right)\sqrt{n} + 128Ka^2\frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}}\ n^{1/4}\right], \tag{19}$$

and

$$T_{p,n} \leq T^*_{p,n} + K^2 \left[1 + \frac{16a\sqrt{\log(2/\delta)}}{\Sigma}\left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right)\sqrt{n} + 128Ka^2\frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}}\ n^{1/4}\right], \tag{20}$$

for any arm $p \leq K$ and any $n \geq \frac{16}{9}c(\delta)^{-2}$, where $c(\delta) = \frac{2a\sqrt{\log(2/\delta)}}{\sqrt{K}(\sqrt{\Sigma}+4a\sqrt{\log(2/\delta)})}$ and $a = 2\sqrt{c_1 \log(c_2/\delta)}$.

Remark. Unlike the bounds for CH-AS in Lemma 4, B-AS allocates the pulls on the arms so that the difference between $T_{p,n}$ and $T^*_{p,n}$ grows with the rate $\tilde{O}(\sqrt{n})$ without dependency on $\lambda_{\min}$. This overcomes the limitation of CH-AS, which as discussed in Remark 3 of Sec. 3.2, may over-sample (thus also under-sample) some arms by $\Omega(n^{2/3})$ whenever $\lambda_{\min}$ is small. We further note that the lower bound in Eq. 19 is of order $\tilde{O}(\lambda_p\sqrt{n})$, which implies that the gap between $T_{p,n}$ and $T^*_{p,n}$ decreases as λ_p becomes smaller. This is not the case in the upper bound, where the gap is of order $\tilde{O}(\sqrt{n})$, but is independent of the value of λ_p. This explains why in the case of general distributions, B-AS has a regret bound with an inverse dependency on $\lambda_{\min}$ (similar to CH-AS), as shown in Thm. 2

Theorem 2. *Assume all the distributions $\{\nu_k\}_{k=1}^K$ are sub-Gaussians with parameters c_1 and c_2. For any $n \geq 0$, the regret of $\mathcal{A}_B$ run with parameters c_1, c_2, and $\delta = n^{-7/2}$ is bounded as*

$$R_n(\mathcal{A}_B) \leq \frac{28230\Big(c_1(c_2+2)^2+1\Big)K^{5/2}\log(n)^2}{\lambda_{\min}n^{3/2}} + \frac{24K}{n^{7/4}\lambda_{\min}}\Big(4368K^2c_1(c_2+2)^2\log(n)^2\Big)^3\Big(\frac{1}{\Sigma^3}+2\Big).$$

Similar to Thm. 1, the bound on the number of pulls translates into a regret bound through Eq. 16. As it can be noticed, in order to remove the dependency on $\lambda_{\min}$, a symmetric bound on $|T_{p,n} - T^*_{p,n}| \leq \tilde{O}(\lambda_p \sqrt{n})$ is needed. While the lower bound in Eq. 19 already decreases with λ_p, the upper bound scales with $\tilde{O}(\sqrt{n})$. Whether there exists an algorithm with a tighter upper bound scaling with λ_p is still an open question. Nonetheless, in the next section, we show that an improved loss bound can be achieved in the special case of Gaussian distributions, which leads to a regret bound without the dependency on $\lambda_{\min}$.

4.3 Regret for Gaussian Distributions

In the case of Gaussian distributions, the loss bound in Eq. 16 can be improved by the following lemma (the full proof is reported in Carpentier et al. 2011).

Lemma 3. *Assume that all the distributions $\{\nu_k\}_{k=1}^K$ are Gaussian. Let $\underline{T}_{k,n}$ be a lower-bound on the number of pulls on an event ξ. Then the loss for arm k satisfies* [8]

$$L_{k,n}(\xi) \leq \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\}\big] \leq \frac{\sigma_k^2}{\underline{T}_{k,n}} \,. \tag{21}$$

Proof (Sketch). A property of Gaussian distributions is that at any time t, the empirical mean $\hat{\mu}_{k,t}$ and variance $\hat{\sigma}^2_{k,t}$ of the rewards of arm k are independent. Since $T_{k,t}$ depends only on $(\hat{\sigma}_{k,t'})_{t'<t}$, it is straightforward to show by induction that $T_{k,n}$ and $\hat{\mu}_{k,n}$ are independent as well. This gives the desired result. □

Remark 1. We notice that the loss bound in Eq. 21 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 $\{\hat{\sigma}_k^2\}_k$ (like in our adaptive algorithms CH-AS and B-AS), we have

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

which might be bigger than $\sigma_k^2/\underline{T}_{k,n}$. In fact, the empirical average $\hat{\mu}_{k,n}$ depends on $T_{k,n}$ through $\{\hat{\sigma}_{k,n}\}_{k=1}^K$ and $\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\} \,|T_{k,n} = t\big]$ is no longer equal to σ_k^2/t. However, Gaussian distributions have the property that the empirical mean $\hat{\mu}_{k,n}$ is independent of the empirical variance $\hat{\sigma}_{k,n}$ (and thus also from $T_{k,n}$), which allows us to obtain the property reported in Lemma 3.

We now report a regret bound in the case of Gaussian distributions. Note that in this case, Assumption 1 holds for $c_1 = 2\Sigma$ and $c_2 = 1$.[9]

[8] The exact definition of the event ξ is available in Sec. B.1 of Carpentier et al. (2011).

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

Theorem 3. *Assume that all distributions $\{\nu_k\}_{k=1}^K$ are Gaussian and that an upper-bound $\overline{\Sigma}$ on Σ is known. The B-AS algorithm run with parameters $c_1 = 2\overline{\Sigma}$, $c_2 = 1$, and $\delta = n^{-7/2}$ has a regret bounded as*

$$R_n(\mathcal{A}_B) \leq \frac{4K}{n^2} + \frac{7060}{n^{3/2}}\Big(c_1(c_2+2)^2+1\Big)K^{3/2}\log(n)^2 + 2560K^2\sqrt{\Sigma(c_2+1)}(c_1(c_2+2)^2\log(n)^2 n^{-7/4}\,. \tag{22}$$

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

Remark 3. 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 b (see Fig. 2). Using the proof of Thm. 3 and the optimized value of δ, it is possible to show that the expected regret is minimized by choosing $b = O\big(\max\{\overline{\Sigma}^{3/2}, \sqrt{\overline{\Sigma}}\}\log n\big)$, which only requires an upper bound on the value of Σ. This is a reasonable assumption whenever a rough estimate of the magnitude of the variances is available.

5 Numerical Experiments

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 algorithms 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 these arm distributions lead to $\lambda_{\min} = 1/5$. Figure 3-*(left)* shows the rescaled regret, $n^{3/2}R_n$, for the three algorithms. Each curve is averaged over $50,000$ runs. The results indicate that while the rescaled regret is almost constant w.r.t. n in B-AS and GAFS-MAX, it increases for small (relative to $\lambda_{\min}^{-1}$) 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 Thm. 2 (Eq. 22). The initial increase in the CH-AS curve is also consistent with the bound of Thm. 1 (Eq. 15). As discussed in Remark 3 of Sec. 3.2, the regret bound for CH-AS is of the form $R_n \leq \min\{\tilde{O}(n^{-3/2}\lambda_{\min}^{-5/2}), \tilde{O}(n^{-4/3})\}$, and thus, the algorithm behaves as $\tilde{O}(n^{-4/3})$ and $\tilde{O}(n^{-3/2}\lambda_{\min}^{-5/2})$ for small and large (relative to $\lambda_{\min}^{-1}$) values of n, respectively. It is important to note that this behavior of CH-AS is independent of the arms' distributions and is intrinsic in the allocation mechanism, as shown in Lemma 4. Finally, the behavior of GAFS-MAX indicates that although its analysis (Antos et al., 2010) shows an inverse dependency on $\lambda_{\min}$ and yields a regret bounds similar to CH-AS, its rescaled regret in fact does not grow with n when the distributions of the arms are Gaussian. This is why we believe that it would be possible to improve the GAFS-MAX analysis by bounding the standard deviation using Bernstein's inequality (e.g., replacing Lemma 2 in Antos et al. 2010 with Lemma 1 in our paper). This would remove the inverse dependency on $\lambda_{\min}$ and provide a regret bound similar to the one for B-AS in the case of Gaussian distributions.

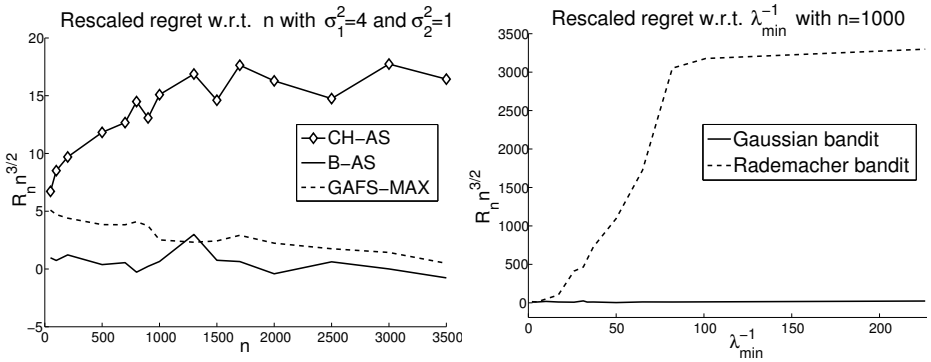

Fig. 3. *(left)* The rescaled regret of CH-AS, B-AS, and GAFS-MAX algorithms on a two-armed problem, where the distributions of the arms are Gaussian. *(right)* The rescaled regret of B-AS for two bandit problems, one with two Gaussian arms and one with a Gaussian and a Rademacher arms.

5.2 B-AS with Non-Gaussian Arms

In Sec. 4.3, we showed that when the arms have Gaussian distributions, the dependency on $\lambda_{\min}$ may be removed from the regret bound of the B-AS algorithm. We also had a discussion on why we conjecture that it is not possible to remove this dependency in case of general distributions unless tighter upper bounds on the number of pulls can be derived. Although we do not yet have a lower bound on the regret showing the dependency on $\lambda_{\min}$, in this section we empirically show that the shape of the distributions has a direct impact on the regret of the B-AS algorithm.

As discussed in Sec. 4.3, the property of Gaussian distributions that allows us to remove the $\lambda_{\min}$ dependency in the regret bound of B-AS is that the empirical mean $\hat{\mu}_{k,n}$ of each arm k is independent of its empirical variance $\hat{\sigma}^2_{k,n}$. Although this property might approximately hold for a larger family of distributions, there are distributions, such as Rademacher, for which these quantities are negatively correlated. In the case of Rademacher distribution,[10] the loss $(\hat{\mu}_{k,t} - \mu_k)^2$ is equal to $\hat{\mu}^2_{k,t}$ and we have $\hat{\sigma}^2_{k,t} = \frac{1}{T_{k,t}} \sum_{i=1}^{t} X^2_{k,i} - \hat{\mu}^2_{k,t} = 1 - \hat{\mu}^2_{k,t}$, as a result, the larger $\hat{\sigma}^2_{k,t}$, the smaller $\hat{\mu}^2_{k,t}$ will be. 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 $\hat{\sigma}^2_{k,t}$, the arm k is pulled more often (given all other things are equal). In case of Rademacher distributions, this means that an arm is over-pulled (pulled more than its optimal allocation) exactly 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 $\nu_2 = \mathcal{N}(0, 1)$,

[10] A random variable X is Rademacher if $X \in \{-1, 1\}$ and admits values -1 and 1 with equal probability.

and one with a Gaussian $\nu_1 = \mathcal{N}(0, \sigma_1^2)$ and a Rademacher $\nu_2 = \mathcal{R}$ 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 $\lambda_{\min}^{-1}$ for $n = 1000$. As expected, while the rescaled regret of B-AS is constant in the first problem, it increases with σ_1^2 in the second one. As explained above, this behavior is due to the poor approximation of the Rademacher arm which is over-pulled whenever its estimated mean is accurate. This result 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-variances (e.g., CH-AS, B-AS, and GAFS-MAX) crucially 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 correlation of these two quantities.

6 Conclusions and Open Questions

In this paper we studied the problem of adaptive allocation for the (uniformly good) estimation of the mean value of K independent distributions first introduced in Antos et al. (2010). Although the algorithm proposed in Antos et al. (2010) achieves a small regret of order $\tilde{O}(n^{-3/2})$, it displays an inverse dependency on the smallest proportion $\lambda_{\min}$. 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 the two algorithms: CH-AS and B-AS. For CH-AS we derived a regret similar to Antos et al. (2010), scaling as $\tilde{O}(n^{-3/2})$ and with the dependence on $\lambda_{\min}$. Unlike in Antos et al. (2010), this result holds for any n and the constants in the bound are made explicit. We then introduced a more refined algorithm, B-AS, which performs an allocation strategy similar to the optimal one. Nonetheless, its general regret bound still depends on $\lambda_{\min}$. 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 a relevant impact on the performance of the allocation strategies.

This work opens a number of questions.

- *Upper bound on the number of pulls.* As mentioned in the Remark of Sec. 4.2, an open question is whether it is possible to devise an allocation algorithm such that $|T_{p,n} - T^*_{p,n}|$ is of order $\tilde{O}(\lambda_p\sqrt{n})$. Such a symmetric bound on the number of pulls would translate into a regret bound without any dependency on $\lambda_{\min}$ for any distribution.
- *Distribution dependency.* Another open question is to which extent the result for B-AS in case of Gaussian distributions could be extended to more general families of distributions. As illustrated in the case of Rademacher, the correlation between the empirical means and variances 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 reward distributions are Gaussian, the empirical means and variances are uncorrelated and the allocation algorithms such as B-AS achieve a better regret. Further investigation is needed to identify whether this results can be extended to other distributions.

– *Lower bound.* The results in Secs. 4.3 and 5.2 suggest that the dependency on the distributions of the arms could be intrinsic in 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 $\lambda_{\min}^{-1}$).

Bibliography

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

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

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.

P. Brémaud. *An Introduction to Probabilistic Modeling*. Springer, 1988.

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.

Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and Peter Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. Technical Report inria-0059413, INRIA, 2011.

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.

P. Chaudhuri and P.A. Mykland. On efficient designing of nonlinear experiments. *Statistica Sinica*, 5:421–440, 1995.

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.

Pierre Étoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12: 335–360, 2010.

V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

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.

A Proof of Theorem 1: The Regret Bound for the CH-AS Algorithm

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. (2010), we begin this section by restating two results from that paper. Let ξ be the event

$$\xi = \bigcap_{1\le k\le K,\ 1\le t\le n} \left\{ \left| \Big(\frac{1}{t}\sum_{i=1}^{t} X_{k,i}^2 - \big(\frac{1}{t}\sum_{i=1}^{t} X_{k,i}\big)^2 \Big) - \sigma_k^2 \right| \le 5\sqrt{\frac{\log(1/\delta)}{2t}} \right\}. \quad (23)$$

Note that the first term in the absolute value in Equation 23 is the sample variance of arm k computed as in Equation 5 for t samples. It can be shown using Hoeffding's inequality that $\Pr(\xi) \ge 1 - 4nK\delta$. 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 a restatement of Lemma 2 in Antos et al. (2010).

Proposition 1. *For any fixed $k = 1, \dots, K$ and $t = 1, \dots, n$, let $\{X_{k,i}\}_i$ be $T_{k,t}$ i.i.d. random variables bounded in $[0, 1]$ from the distribution ν_k with mean μ_k and 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:*

$$|\hat{\sigma}_{k,t}^2 - \sigma_k^2| \le 5\sqrt{\frac{\log(1/\delta)}{2T_{k,t}}}\ . \quad (24)$$

We also need to draw a connection between the allocation and stopping time problems. Thus, we report the following proposition which is a restatement of Lemma 10 in Antos et al. (2010).

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

$$\mathbb{E}\left[\Big(\sum_{i=1}^{T} X_i - T\,\mu \Big)^2 \right] \le \mathbb{E}[T]\ \sigma^2. \quad (25)$$

A.2 Allocation Performance and Regret Bound

In this section, we first provide the proof of Lemma 4 and then use the result to prove Theorem 1.

Lemma 4. *Assume that the supports of the distributions $\{\nu_k\}_{k=1}^K$ are in $[0,1]$ and that $n \geq 4K$. For any $\delta > 0$, for any arm $1 \leq k \leq K$, the number of pulls $T_{k,n}$ played by the CH-AS algorithm satisfies with probability at least $1 - 4nK\delta$,*

$$-\frac{5}{\Sigma^2 \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} - \frac{K}{\Sigma} \leq T_{k,n} - T_{k,n}^* \leq \frac{5(K-1)}{\Sigma^2 \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} + \frac{K^2}{\Sigma} \ . \quad (26)$$

Proof (Lemma 4). The proof consists of the following three main steps.

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

$$B_{q,t+1} = \frac{1}{T_{q,t}} \left(\hat{\sigma}_{q,t}^2 + 5\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}} \right), \qquad 1 \leq q \leq K \ .$$

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

$$\frac{\sigma_q^2}{T_{q,t}} \leq B_{q,t+1} \leq \frac{1}{T_{q,t}} \left(\sigma_q^2 + 10\sqrt{\frac{\log(1/\delta)}{2T_{q,t}}} \right). \quad (27)$$

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

$$B_{p,t+1} \leq B_{k,t+1} \ . \quad (28)$$

From Equation 27 and the fact that $T_{k,t} = T_{k,n} - 1$, we obtain

$$B_{k,t+1} \leq \frac{1}{T_{k,t}} \left(\sigma_k^2 + 10\sqrt{\frac{\log(1/\delta)}{2T_{k,t}}} \right) = \frac{1}{T_{k,n}-1} \left(\sigma_k^2 + 10\sqrt{\frac{\log(1/\delta)}{2(T_{k,n}-1)}} \right). \quad (29)$$

Using the lower bound in Equation 27 and the fact that $T_{p,t} \leq T_{p,n}$, we may lower bound $B_{p,t}$ as

$$B_{p,t+1} \geq \frac{\sigma_p^2}{T_{p,t}} \geq \frac{\sigma_p^2}{T_{p,n}} \ . \quad (30)$$

Combining Equations 28, 29, and 30, we obtain

$$\frac{\sigma_p^2}{T_{p,n}} \leq \frac{1}{T_{k,n}-1} \left(\sigma_k^2 + 10\sqrt{\frac{\log(1/\delta)}{2(T_{k,n}-1)}} \right) . \quad (31)$$

Note that at this point there is no dependency on t, and thus, Equation 31 holds with probability $1 - 4nK\delta$ (this is because Equation 31 is defined on the event ξ)

for an 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 p is under-pulled *without taking into account the initialization phase*, i.e., $T_{p,n} - 2 < \lambda_p(n - 2K)$, then from the constraint $\sum_k (T_{k,n} - 2) = n - 2K$, 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 Eq. 11. Using the definition of the optimal allocation $T^*_{k,n} = n\lambda_k = n\sigma_k^2/\Sigma$ and the fact that $T_{k,n} \geq \lambda_k(n - 2K) + 2$, Eq. 11 may be written as

$$\frac{\sigma_p^2}{T_{p,n}} \leq \frac{1}{T^*_{k,n}} \frac{n}{n-2K} \left(\sigma_k^2 + \sqrt{\frac{100 \log(1/\delta)}{2(\lambda_k(n-2K)+2-1)}} \right) \leq \frac{\Sigma}{n} + \frac{20\sqrt{\log(1/\delta)}}{(\lambda_{\min} n)^{3/2}} + \frac{4K\Sigma}{n^2}, \tag{32}$$

since $\lambda_k(n - 2K) + 1 \geq \lambda_k(n/2 - 2K + 2K) + 1 \geq \frac{n\lambda_k}{2}$, as $n \geq 4K$ (thus also $\frac{2K\Sigma}{n(n-2K)} \leq \frac{4K\Sigma}{n^2}$). By reordering the terms in the previous equation, we obtain the lower bound

$$T_{p,n} \geq \frac{\sigma_p^2}{\frac{\Sigma}{n} + \frac{20\sqrt{\log(1/\delta)}}{(n\lambda_{\min})^{3/2}} + \frac{4K\Sigma}{n^2}} \geq T^*_{p,n} - \frac{5\sqrt{n \log(1/\delta)}}{\Sigma^2 \lambda_{\min}^{3/2}} - \frac{K}{\Sigma}, \tag{33}$$

where in the second inequality we used $1/(1 + x) \geq 1 - x$ (for $x > -1$) and $\sigma_p^2 \leq 1/4$. Note that the lower bound holds w.h.p. for any arm p.

Step 3. Upper bound on $T_{p,n}$. Using Eq. 33 and the fact that $\sum_k T_{k,n} = n$, we obtain the upper bound

$$T_{p,n} = n - \sum_{k \neq p} T_{k,n} \leq T^*_{p,n} + \frac{5(K-1)}{\Sigma^2 \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} + \frac{K^2}{\Sigma} . \tag{34}$$

The claim follows by combining the lower and upper bounds in Equations 33 and 34.

□

We are now ready to prove Theorem 1.

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

Step 1. $T_{k,n}$ is a stopping time. For each arm $1 \leq k \leq K$, let $\{X_{k,t}\}_{t \leq n}$ be all the samples collected from pulling that arm. We first show that $T_{k,n}$ is a stopping time adapted to the process $(X_{k,t})_{t \leq n}$. At each time step t, the CH-AS algorithm decides which arm to pull only according to the current values

of the upper-bounds $\{B_{k,t}\}_k$. Thus for any arm k, $T_{k,(t+1)}$ depends only on the values $\{T_{k,t}\}_k$ and $\{\hat{\sigma}^2_{k,t}\}_k$. So by induction, $T_{k,(t+1)}$ depends on the sequence $\{X_{k,1}, \ldots, X_{k,T_{k,t}}\}$, and on the realizations of the other arms (whose randomness is independent of the value of arm k), and thus, we may conclude that $T_{k,n}$ is a stopping time adapted to the process $(X_{k,t})_{t \leq n}$.

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 we bound the first term as

$$\mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\}\Big] \leq \frac{\sigma_k^2 \mathbb{E}[T_{k,n}]}{\underline{T}^2_{k,n}} \ , \tag{35}$$

where $\underline{T}_{k,n}$ is the lower bound on $T_{k,n}$ on the event ξ. Since the upper-bound in Lemma 4 is obtained with high probability with respect to the event ξ, we may easily convert it to a bound in expectation as follows:

$$\mathbb{E}[T_{k,n}] \leq \left(T^*_{p,n} + \frac{5(K-1)}{\Sigma^2 \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} + \frac{K^2}{\Sigma} \right) + n \times 4nK\delta. \tag{36}$$

Combining Equation 35 and 36, and using Equation 32 for $\sigma_k^2 / \underline{T}_{k,n}$, we obtain

$$\begin{aligned} &\mathbb{E}\Big[(\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\}\Big] \\ &\leq \left(\frac{\Sigma}{n} + \frac{20\sqrt{\log(1/\delta)}}{(\lambda_{\min} n)^{3/2}} + \frac{4K\Sigma}{n^2} \right)^2 \frac{\left(T^*_{p,n} + \frac{5(K-1)}{\Sigma^2 \lambda_{\min}^{3/2}} \sqrt{n \log(1/\delta)} + \frac{K^2}{\Sigma} + n \times 4nK\delta \right)}{\sigma_k^2}. \end{aligned} \tag{37}$$

By setting $A = \frac{5\sqrt{\log(1/\delta)}}{\lambda_{\min}^{3/2}}$ to simplify the notation, Equation 37 may be rewritten as

$$
\begin{aligned}
&\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\Big]\\
&\le \left(\frac{\Sigma^2}{n^2}+\frac{2\Sigma}{n^2}\Big(\frac{4A}{\sqrt{n}}+\frac{4K\Sigma}{n}\Big)+\frac{32A^2+32\Sigma^2K^2}{n^3}\right)\left(\frac{n}{\Sigma}+\frac{AK}{\Sigma^2\sigma_k^2}\sqrt{n}+\frac{K^2}{\Sigma}+4n^2K\delta\right)\\
&\le \frac{\Sigma^2}{n^2}\left(\frac{n}{\Sigma}+\frac{AK}{\Sigma^2\sigma_k^2}\sqrt{n}+\frac{K^2}{\Sigma}+4n^2K\delta\right)\\
&+\left(\frac{8A\Sigma}{n^{5/2}}+\frac{32A^2+40\Sigma^2K^2}{n^3}\right)\left(\frac{n}{\Sigma}+\frac{AK^2}{\sigma_k^2}\sqrt{n}(\frac{1}{\Sigma^2}+\frac{1}{\Sigma})+4n^2K\delta\right)\\
&\le \frac{\Sigma}{n}+\frac{AK}{\sigma_k^2n^{3/2}}+\frac{\Sigma K^2}{n^2}+4\Sigma^2K\delta+\frac{8A}{n^{3/2}}+\frac{8A^2K^2}{\sigma_k^2n^2}(\frac{1}{\Sigma}+1)+\frac{32AK\Sigma\delta}{\sqrt{n}}\\
&+8\frac{4A^2+5\Sigma^2K^2}{n^2\Sigma}+8AK^2\frac{4A^2+5\Sigma^2K^2}{\sigma_k^2n^{5/2}}(\frac{1}{\Sigma^2}+\frac{1}{\Sigma})+32K\delta\frac{4A^2+5\Sigma^2K^2}{\Sigma n}\\
&\le \frac{\Sigma}{n}+\frac{A}{n^{3/2}}(\frac{K}{\sigma_k^2}+8)+\frac{K^2}{n^2}\Big(41\Sigma+\frac{8A^2}{\sigma_k^2}(5+1/\Sigma)\Big)\\
&+8AK^2\frac{4A^2+5\Sigma^2K^2}{\sigma_k^2n^{5/2}}(\frac{1}{\Sigma^2}+\frac{1}{\Sigma})+4K^3A^2\delta(15\Sigma^2+\frac{22}{\Sigma})\delta\\
&\le \frac{\Sigma}{n}+\frac{A}{n^{3/2}}(\frac{K}{\sigma_k^2}+8)+\frac{K^2}{n^2}\Big(41\Sigma+\frac{128A^3K^2}{\sigma_k^2}(\frac{16}{\Sigma^2}+16\Sigma)\Big)\\
&+4K^3A^2\delta(15\Sigma^2+\frac{22}{\Sigma})\delta.
\end{aligned}
$$

Note that if $\delta = n^{-5/2}$, we have $A \le \frac{8\log(n)}{\lambda_{\min}^{3/2}}$ and also, we obtain

$$
\begin{aligned}
&\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi\}\Big]\\
&\le \frac{\Sigma}{n}+\frac{A}{n^{3/2}}(\frac{K}{\sigma_k^2}+8+4\Sigma^2K)+\frac{K^2}{n^2}\Big(41\Sigma+\frac{128A^3K^2}{\sigma_k^2}(\frac{18}{\Sigma^2}+18\Sigma^2)\Big)\\
&\le \frac{\Sigma}{n}+\frac{8\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}(\frac{K}{\lambda_{\min}\Sigma^2}+8+4\Sigma^2K)+\frac{65536\log(n)^3K^4}{n^2\lambda_{\min}^7\Sigma^2}(\frac{18}{\Sigma^2}+18\Sigma^2).
\end{aligned}
\tag{38}
$$

Since $|\hat{\mu}_{k,n}-\mu_k|$ is always smaller than 1, we have $\mathbb{E}\Big[(\hat{\mu}_{k,n}-\mu_k)^2\mathbb{I}\{\xi^C\}\Big] \le 4nK\delta \le \frac{4K}{n^{3/2}}$. Thus using Equation 38, the expected loss of arm k is bounded by

$$
\begin{aligned}
L_{k,n} &\le \frac{\Sigma}{n}+\frac{8\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}(\frac{K}{\lambda_{\min}\Sigma^2}+8+4\Sigma^2K)+\frac{65536\log(n)^3K^4}{n^2\lambda_{\min}^7\Sigma^2}(\frac{18}{\Sigma^2}+18\Sigma^2)+4nK\delta\\
&\le \frac{\Sigma}{n}+\frac{8\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}(\frac{K}{\lambda_{\min}\Sigma^2}+8K+4\Sigma^2K)+\frac{65536\log(n)^3K^4}{n^2\lambda_{\min}^7\Sigma^2}(\frac{18}{\Sigma^2}+18\Sigma^2).
\end{aligned}
$$

Using the definition of regret $R_n = \max_k L_{k,n} - \frac{\Sigma}{n}$, we obtain

$$R_n \leq \frac{8\log(n)}{n^{3/2}\lambda_{\min}^{3/2}}\left(\frac{K}{\lambda_{\min}\Sigma^2} + 8K + 4\Sigma^2 K\right) + \frac{65536\log(n)^3 K^4}{n^2\lambda_{\min}^7\Sigma^2}\left(\frac{18}{\Sigma^2} + 18\Sigma^2\right). \quad (39)$$

□

A.3 Example of an Alternative Regret Bound

We report a sketch of the proof for the example with $\lambda_{\min}$ reported in the Remark 3 of Section 3.2. Using the definition of $B_{k,t+1}$ and Proposition 1, since $\hat{\sigma}_{2,t}^2 = 0$, we have that at any time $t+1 > 4$

$$B_{1,t+1} \leq \frac{1}{T_{1,t}}\Big(1 + 10\sqrt{\frac{\log(1/\delta)}{2}}\Big) \qquad \text{and} \qquad B_{2,t+1} = \frac{1}{T_{2,t}}\Big(5\sqrt{\frac{\log(1/\delta)}{2T_{2,t}}}\Big). \quad (40)$$

Let $t+1 \leq n$ be the last time that arm 1 was pulled, i.e., $T_{1,t} = T_{1,n} - 1$ and $B_{1,t+1} \geq B_{2,t+1}$. From Equation 40, we have

$$B_{2,t+1} = \frac{1}{T_{2,t}}\Big(5\sqrt{\frac{\log(1/\delta)}{2T_{2,t}}}\Big) \leq B_{1,t+1} \leq \frac{1}{T_{1,n}-1}\Big(1 + 10\sqrt{\frac{\log(1/\delta)}{2}}\Big). \quad (41)$$

Now consider the two possible cases: **1)** $T_{1,n} \leq n/2$, in which case obviously $T_{2,n} \geq n/2$ and **2)** $T_{1,n} > n/2$, in this case Equation 41 implies that $T_{2,n} \geq T_{2,t} = \widetilde{\Omega}\big(n^{2/3}\big)$. Thus in both cases, we may write $T_{2,n} = \widetilde{\Omega}\big(n^{2/3}\big)$, which indicates that arm 2 (resp. arm 1) is over-sampled (resp. under-sampled) by a number of pulls of order $\widetilde{O}(n^{2/3})$. By following the same arguments as in the proof of Theorem 1, we deduce that the regret in this case is of order $\widetilde{O}(n^{-4/3})$.

B Proof of Theorems 2 and 3: The Regret Bounds for the Bernstein Algorithm

B.1 Basic Tools

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

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

Lemma 5. *Let Assumption 1 hold and $n \geq 2$. Define the following event*

$$\xi = \xi_{K,n}(\delta) = \bigcap_{1\leq k\leq K,\ 2\leq t\leq n} \left\{ \left| \sqrt{\frac{1}{t-1}\sum_{i=1}^{t}\left(X_{k,i} - \frac{1}{t}\sum_{j=1}^{t} X_{k,j}\right)^2} - \sigma_k \right| \leq 2a\sqrt{\frac{\log(2/\delta)}{t}} \right\}, \tag{42}$$

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

Note that the first term in the absolute value in Equation 42 is the empirical standard deviation of arm k computed as in Equation 17 for t samples. The event ξ plays an important role in the proofs of this section and a number of statements will be proved on this event.

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

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

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

$$\begin{aligned}
\Big|\mathbb{E}\Big[|X_{k,t} - \mu_k|^2 \ \mid\ |X_{k,t} - \mu_k|^2 \leq b\Big] - \sigma_k^2\Big| &= \frac{\Big|\mathbb{E}\Big[\Big((X_{k,t} - \mu_k)^2 - \sigma_k^2\Big)\mathbb{I}\left\{|X_{k,t} - \mu_k|^2 > b\right\}\Big]\Big|}{\mathbb{P}\Big(|X_{k,t} - \mu_k|^2 \leq b\Big)} \\
&\leq \frac{c_1\delta(1 + \log(c_2/\delta)) + \delta\sigma_k^2}{1-\delta}.
\end{aligned} \tag{43}$$

Note also that Cauchy-Schwartz inequality implies

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

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

$$|\tilde{\mu}_k-\mu_k|=\frac{\Big|\mathbb{E}\Big[\Big(X_{k,t}-\mu_k\Big)\mathbb{I}\left\{|X_{k,t}-\mu_k|^2>b\right\}\Big]\Big|}{\mathbb{P}\Big(|X_{k,t}-\mu_k|^2\leq b\Big)}\leq\frac{\sqrt{c_1\delta(1+\log(c_2/\delta))}}{1-\delta}, \tag{44}$$

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

We note $\tilde{\sigma}_k^2=\mathbb{V}\Big[X_{k,t}\ \mid\ |X_{k,t}-\mu_k|^2\leq b\Big]=\mathbb{E}\Big[|X_{k,t}-\mu_k|^2\ \mid\ |X_{k,t}-\mu_k|^2\leq b\Big]-(\mu_k-\tilde{\mu_k})^2$. From Equations 43 and 44, we derive

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

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

$$|\tilde{\sigma_k}-\sigma_k|\leq\frac{\sqrt{2c_1\delta(1+c_2+\log(c_2/\delta))}}{1-\delta}. \tag{45}$$

Step 2. Application of large deviation inequalities.

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

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

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

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

Theorem 4 (Maurer and Pontil (2009)). *Let $(X_1,...,X_t)$ be $t\geq 2$ i.i.d. random variables of variance σ^2 and mean μ and such that $\forall i\leq t, X_i\in[a,a+c]$. Then with probability at least $1-\delta$:*

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

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

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

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

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

Now, from Equation 45, we have on $\xi_1 \cap \xi_2$, for all $1 \le k \le K,\ 2 \le t \le n$:

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

from which we deduce Lemma 5 (since $\xi_1 \cap \xi_2 \subseteq \xi$ and $2 \le t \le n$).

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

Corollary 1. *For any $k = 1, \dots, K$ and $t = 2K, \dots, n$, let $\{X_{k,i}\}_i$ be n i.i.d. random variables drawn from ν_k, satisfying Assumption 1. Let $T_{k,t}$ be any random variable taking values in $\{2, \dots, n\}$. Let $\hat{\sigma}^2_{k,t}$ be the empirical variance computed from Equation 17. Then, on the event ξ, we have:*

$$|\hat{\sigma}_{k,t} - \sigma_k| \le 2a\sqrt{\frac{\log(2/\delta)}{T_{k,t}}}\ . \tag{46}$$

Other important properties

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

Proposition 3. *Let $\{\mathcal{F}_t\}$ be a filtration and X_t a $\mathcal{F}_t$-adapted sequence of i.i.d. random variables with variance σ^2. Assume that $\mathcal{F}_t$ and the σ-algebra generated by $\{X_i : i \ge t+1\}$ are independent and T is a stopping time w.r.t. $\mathcal{F}_t$ with a finite expected value. If $\mathbb{E}[X_1^2] < \infty$ then*

$$\mathbb{E}\left[\Big(\sum_{i=1}^{T} X_i - T\,\mu\Big)^2\right] = \mathbb{E}[T]\ \sigma^2. \tag{47}$$

Bound on the regret outside of ξ. The next lemma provides a bound for the loss whenever the event ξ does not hold.

Lemma 6. *Let Assumption 1 holds. Then for every arm k:*

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

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

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

and thus by setting $\epsilon = c_1 \log(c_2/2nK\delta)$[11], we obtain

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

We thus know that

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

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

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

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

Technical inequalities

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

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

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

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

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

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

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

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

We then have

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

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

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

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

Upper bound on $\mathbb{E}\big[|\hat{\mu}_{k,n} - \mu_k|^2 \mathbb{I}\left\{\xi^C\right\}\big]$ when $\delta = n^{-7/2}$: We get from Lemma 6 when $\delta = n^{-7/2}$ and when choosing c_2 such that $c_2 \geq 2enK\delta = 2Ken^{-5/2}$

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

Upper bound on B for $\delta = n^{-7/2}$

$$
\begin{aligned}
B &= 16Ka\sqrt{\log(2/\delta)}\left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right) \\
&\leq 36K\sqrt{\log(n)}a\Big(\sqrt{\Sigma} + 12\sqrt{K}(\sqrt{c_2}+1)a\sqrt{\log(n)}\Big) \\
&\leq 36K\sqrt{\log(n)}a\sqrt{\Sigma} + 432K^{3/2}(\sqrt{c_2}+1)a^2\log(n) \\
&\leq 72\sqrt{2}K\log(n)\sqrt{c_1(c_2+1)}\sqrt{\Sigma} + 3456c_1(c_2+2)K^{3/2}(\sqrt{c_2}+1)\log(n)^2 \\
&\leq 3559\Big(c_1(c_2+2)^2+1\Big)K^{3/2}\log(n)^2.
\end{aligned}
$$

Upper bound on C for $\delta = n^{-7/2}$

$$
\begin{aligned}
C &= 64\sqrt{2}K^{3/2}a^2\frac{\log(2/\delta)}{\sqrt{c(\delta)}} \\
&\leq 640K^2(\sqrt{c_2}+1)\log(n)a^2 \\
&\leq 640K^2c_1(c_2+2)^2\log(n)^2 \\
&\leq 640K^2c_1(c_2+2)^2\log(n)^2.
\end{aligned}
$$

B.2 Allocation Performance and Regret Bounds for Sub-Gaussian Distributions

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 the Theorem 3 for Gaussians.

Proof (Lemma 2). The proof consists of the following five main steps.

Step 1. Lower bound of order $O(\sqrt{n})$. Let k be the index of an arm such that $T_{k,n} \geq \frac{n}{K}$ and $t+1 \leq n$ be the last time that it was pulled, i.e., $T_{k,t} = T_{k,n} - 1$ and $T_{k,t+1} = T_{k,n}$. [12] From Equation 46 and the fact that $T_{k,n} \geq \frac{n}{K} \geq 4$, we obtain on ξ

$$
B_{k,t+1} \leq \frac{1}{T_{k,t}}\left(\sigma_k + 4a\sqrt{\frac{\log(2/\delta)}{T_{k,t}}}\right)^2 \leq \frac{K\Big(\sqrt{\Sigma} + 4a\sqrt{\log(2/\delta)}\Big)^2}{n}, \tag{48}
$$

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

$$
B_{q,t+1} \leq B_{k,t+1}. \tag{49}
$$

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

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

$$B_{q,t+1} \geq 4a^2 \frac{\log(2/\delta)}{T_{q,t}^2} \geq 4a^2 \frac{\log(2/\delta)}{T_{q,n}^2} \ . \tag{50}$$

Combining Equations 48–50, we obtain

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

Finally, this implies that for any q

$$T_{q,n} \geq \frac{2a\sqrt{\log(2/\delta)}}{\left(\sqrt{\Sigma} + 4a\sqrt{\log(2/\delta)}\right)} \sqrt{\frac{n}{K}}. \tag{51}$$

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

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

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

Step 2. Mechanism of the algorithm. Similar to Step 1 of the proof of Lemma 4, we first recall 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 1 it follows that on ξ, for any q,

$$\frac{\sigma_q^2}{T_{q,t}} \leq B_{q,t+1} \leq \frac{1}{T_{q,t}} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right)^2. \tag{52}$$

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

$$B_{p,t+1} \leq B_{q,t+1} \ . \tag{53}$$

From Equation 52 and $T_{q,t} = T_{q,n} - 1$, we obtain

$$B_{q,t+1} \leq \frac{1}{T_{q,t}} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,t}}} \right)^2 = \frac{1}{T_{q,n}-1} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}} \right)^2. \tag{54}$$

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

$$B_{p,t+1} \geq \frac{\sigma_p^2}{T_{p,t}} \geq \frac{\sigma_p^2}{T_{p,n}}. \tag{55}$$

Combining Equations 53–55, we obtain

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

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

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \leq \sum_{q|T_{q,n}>2} \left(\sigma_q + 4a\sqrt{\frac{\log(2/\delta)}{T_{q,n}-1}}\right)^2, \tag{56}$$

because the arms that are not pulled after the initialization are only pulled twice.
Step 3. Intermediate lower bound. It is possible to rewrite Equation 56, using the fact that $T_{q,n} \geq 2$, as

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

Plugging Equation 51 in Equation 56, we have on ξ for any arm p

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \leq \sum_q \left(\sigma_q + 4a\sqrt{\frac{2\log(2/\delta)}{T_{q,n}-1}}\right)^2 \leq \left(\sqrt{\Sigma} + 4\sqrt{K}a\sqrt{2\frac{\log(2/\delta)}{c(\delta)\sqrt{n}}}\right)^2, \tag{57}$$

because for any sequence $(a_k)_{i=1,\dots,K} \geq 0$, and any $b \geq 0$, $\sum_k (a_k+b)^2 \leq (\sqrt{\sum_k a_k^2} + \sqrt{K}b)^2$ by Cauchy Schwartz.
Building on this bound we may finally recover the desired bound.

Step 4. Final lower bound. We first develop the square in Equation 56 using $T_{q,n} \geq 2$ as

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

We now use the bound in Equation 57 in the second term of the RHS and the bound in Equation 51 to bound $T_{k,n}$ in the last term, thus obtaining

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

By using again $n \geq 4K$ and some algebra, we get

$$\begin{aligned}\frac{\sigma_p^2}{T_{p,n}}(n-2K) &\leq \Sigma + 16Ka\sqrt{\frac{\Sigma\log(2/\delta)}{n}} + 64\sqrt{2}K^{3/2}a^2\frac{\log(2/\delta)}{\sqrt{c(\delta)}}n^{-3/4} + \frac{32Ka^2\log(2/\delta)}{c(\delta)\sqrt{n}} \\ &= \Sigma + \frac{16Ka\sqrt{\log(2/\delta)}}{\sqrt{n}}\left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right) + 64\sqrt{2}K^{3/2}a^2\frac{\log(2/\delta)}{\sqrt{c(\delta)}}n^{-3/4}.\end{aligned} \tag{58}$$

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

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

Note that the above lower bound holds with high probability 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

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

□

We can now prove a general bound for the regret of this algorithm.

Proof (Theorem 2).

At first let us call, for the sake of convinience,

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

Then Equation 58 easily becomes

$$\frac{\sigma_p^2}{T_{p,n}}(n-2K) \leq \Sigma + \frac{B}{\sqrt{n}} + \frac{C}{n^{3/4}}. \tag{59}$$

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

$$T_{p,n} \leq T^*_{p,n} + \frac{BK}{\Sigma}\sqrt{n} + \frac{CK}{\Sigma}n^{1/4} + 2K.$$

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

$$\mathbb{E}(T_{p,n}) \leq T^*_{p,n} + \frac{BK}{\Sigma}\sqrt{n} + \frac{CK}{\Sigma}n^{1/4} + 2K + n \times 4nK\delta \tag{60}$$

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

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

By combining that and Equations 59, 60, and 47 (as done in Equation 35), we obtain for the first part of the loss and because:

$$\begin{aligned}
&\mathbb{E}[(\hat{\mu}_{k,n}-\mu)^2\mathbb{I}\{\xi\}] \\
&\leq \frac{1}{\sigma_p^2(n-2K)^2}\Big(\Sigma + \frac{B}{\sqrt{n}} + \frac{C}{n^{3/4}}\Big)^2\Big(T^*_{p,n} + \frac{KB}{\Sigma}\sqrt{n} + \frac{CK}{\Sigma}n^{1/4} + 2K + 4n^2K\delta\Big) \\
&\leq \frac{1}{(n-2K)^2}\left(\Sigma^2 + 2\Sigma(\frac{B}{\sqrt{n}} + \frac{C}{n^{3/4}}) + \frac{(B+C)^2}{n^{3/2}}\right)\Big(\frac{n}{\Sigma} + \frac{KB}{\Sigma^2\lambda_k}\sqrt{n} + \frac{CK}{\Sigma^2\lambda_k}n^{1/4} + \frac{2K}{\Sigma\lambda_k} + \frac{4n^2K\delta}{\Sigma\lambda_k}\Big) \\
&\leq \frac{1}{(n-2K)^2}\Bigg(n\Sigma + \frac{KB}{\lambda_k}\sqrt{n} + \frac{CK+2K\Sigma}{\lambda_k}n^{1/4} + \frac{4n^2K\Sigma\delta}{\lambda_k} + 2B\sqrt{n} + 2Cn^{1/4} \\
&+ \frac{2(B+C)(\frac{KB}{\Sigma} + \frac{KC}{\Sigma} + 2K)}{\lambda_k} + \frac{8(B+C)n^2K\delta}{\lambda_k} + \frac{(B+C)^2}{\Sigma\lambda_k\sqrt{n}}\Big(\frac{K(B+C)}{\Sigma} + 3K\Big) + 4n^2K\delta\frac{(B+C)^2}{\Sigma\lambda_k}\Bigg) \\
&\leq \frac{1}{(n-2K)^2}\Bigg(n\Sigma + \frac{2KB}{\lambda_k}\sqrt{n} + \frac{K}{\lambda_k}\Big[\Big(\frac{(B+C)}{\Sigma} + 3\Big)^3 + 2(C+\Sigma)\Big]n^{1/4} \\
&+ \frac{4n^2K\delta}{\lambda_k}\Big(2(B+C) + \Sigma + (B+C)^2\Big)\Bigg).
\end{aligned}$$

Now note, as $\delta = n^{-5/2}$, that

$$
\begin{aligned}
&\mathbb{E}[(\hat{\mu}_{k,n}-\mu)^2\mathbb{I}\{\xi\}] \\
&\leq \frac{\Sigma}{n}+\frac{8KB}{\lambda_k n^{3/2}}+\frac{4K}{n^{7/4}\lambda_k}\Big[2\Big(\frac{(B+C)}{\Sigma}+3\Big)^3+2(C+\Sigma)\Big] \\
&+\frac{16Kn^{-5/2}}{\lambda_k}\Big(2(B+C)+\Sigma+(B+C)^2)\Big) \\
&\leq \frac{\Sigma}{n}+\frac{8KB}{\lambda_k n^{3/2}}+\frac{24K}{n^{7/4}\lambda_k}\Big(\frac{(B+C)}{\Sigma}+3\Big)^3\Big(1+2\Sigma^3\Big).
\end{aligned}
$$

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

$$
R_n \leq \frac{8KB}{\lambda_{\min} n^{3/2}}+\frac{24K}{n^{7/4}\lambda_{\min}}\Big(\frac{(B+C)}{\Sigma}+3\Big)^3\Big(1+2\Sigma^3\Big)+2c_1 nK\delta(1+\log(c_2/2nK\delta)).
$$

By recalling the bounds on B and C in Appendix B.1 and taking $\delta = n^{-5/2}$, we obtain:

$$
\begin{aligned}
R_n &\leq \frac{8KB}{\lambda_{\min} n^{3/2}}+\frac{24K}{n^{7/4}\lambda_{\min}}\Big(\frac{(B+C)}{\Sigma}+3\Big)^3\Big(1+2\Sigma^3\Big)+6c_1c_2K\log(n)n^{-3/2} \\
&\leq \frac{28230\Big(c_1(c_2+2)^2+1\Big)K^{5/2}\log(n)^2}{\lambda_{\min} n^{3/2}} \\
&+\frac{24K}{n^{7/4}\lambda_{\min}}\Big(4368K^2c_1(c_2+2)^2\log(n)^2\Big)^3\Big(\frac{1}{\Sigma^3}+2\Big).
\end{aligned}
$$

□

B.3 Regret Bound for Gaussian Distributions

Here we report the proof of Lemma 3 which states 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). Before reporting the proof of Lemma 3, we recall a property of the normal distribution that is used in this proof (see e.g., Brémaud 1988).

Proposition 4. *Let $X_1,\dots,X_n$ be n i.i.d. Gaussian random variables. Then their empirical mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and empirical variance $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu}_n)^2$ are independent of each other.*

Proof (Lemma 3).

Let $T \geq 2$ be a integer-valued random variable, which is a stopping time with respect to the filtration $\mathcal{F}_t$ generated by the sequence of i.i.d. random variables X_t drawn from a Gaussian distribution. Write $\hat{\mu}_t = \frac{1}{t}\sum_{i=1}^t X_i$ and $\hat{\sigma}_t^2 = \frac{1}{t-1}\sum_{i=1}^t (X_i - \hat{\mu}_t)^2$ the empirical mean and variance of the t first samples.

Lemma 7. *We have*

$$\hat{\sigma}_{t+1}^2 = \frac{t-1}{t}\hat{\sigma}_t^2 + \frac{1}{t+1}(X_{t+1} - \hat{\mu}_t)^2,$$

which leads to for any $t \geq 3$ to

$$\hat{\sigma}_t^2 = a_{2,t}\sigma_2^2 + \sum_{i=3}^{t} a_{i,t}(X_i - \hat{\mu}_{i-1})^2,$$

where the $(a_{i,t})_i$ are real numbers.

Proof. At first note that from the definition of empirical variance and empirical mean

$$\begin{aligned}
\hat{\sigma}_{t+1}^2 &= \frac{1}{t}\sum_{i=1}^{t+1}(X_i - \hat{\mu}_{t+1})^2 \\
&= \frac{1}{t}\sum_{i=1}^{t}(X_i - \hat{\mu}_{t+1} + \hat{\mu}_t - \hat{\mu}_t)^2 + \frac{1}{t}(X_{t+1} - \hat{\mu}_{t+1})^2 \\
&= \frac{1}{t}\sum_{i=1}^{t}(X_i - \hat{\mu}_t)^2 + \frac{1}{t}(X_{t+1} - \hat{\mu}_{t+1})^2 + (\hat{\mu}_t - \hat{\mu}_{+1})^2 \\
&= \frac{1}{t}\sum_{i=1}^{t}(X_i - \hat{\mu}_t)^2 + \frac{t}{(t+1)^2}(X_{t+1} - \hat{\mu}_t)^2 + \frac{1}{(t+1)^2}(X_{t+1} - \hat{\mu}_t)^2 \\
&= \frac{1}{t}\sum_{i=1}^{t}(X_i - \hat{\mu}_t)^2 + \frac{t}{t+1}(X_{t+1} - \hat{\mu}_t)^2,
\end{aligned}$$

which proofs the first part of the lemma.
We now want to prove the second part by induction.
Note that for $t = 3$, the property we just proved gives us

$$\hat{\sigma}_3^2 = \frac{t-1}{t}\hat{\sigma}_2^2 + \frac{1}{3}(X_3 - \hat{\mu}_2)^2,$$

and the property is true for $t = 3$.
Now assume that for a given $t \geq 3$, we have

$$\hat{\sigma}_t^2 = a_{2,t}\sigma_2^2 + \sum_{i=3}^{t} a_{i,t}(X_i - \hat{\mu}_{i-1})^2,$$

where the $(a_{i,t})_i$ are real numbers.
Note that by using $\hat{\sigma}_{t+1}^2 = \frac{t-1}{t}\hat{\sigma}_t^2 + \frac{1}{t+1}(X_{t+1} - \hat{\mu}_t)^2$, we get

$$\hat{\sigma}_{t+1}^2 = \frac{t-1}{t}\Big(a_{2,t}\sigma_2^2 + \sum_{i=3}^{t} a_{i,t}(X_i - \hat{\mu}_{i-1})^2\Big) + \frac{1}{t+1}(X_{t+1} - \hat{\mu}_t)^2$$
$$= \frac{t-1}{t}a_{2,t}\sigma_2^2 + \sum_{i=3}^{t} \frac{t-1}{t}a_{i,t}(X_i - \hat{\mu}_{i-1})^2\Big) + \frac{1}{t+1}(X_{t+1} - \hat{\mu}_t)^2,$$

and then the property holds at time $t+1$ with $a_{2,t+1} = \frac{t-1}{t}a_{2,t}$ and $\forall 2 \leq i \leq t$, $a_{i,t+1} = \frac{t-1}{t}a_{i,t}$ and $a_{t+1,t+1} = \frac{1}{t+1}$.

Lemma 8.

$$\forall 2 \leq t \leq n, \qquad \hat{\mu}_t \Big| \Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \leq t-1}\Big\} \sim \mathcal{N}(\mu, \frac{\sigma^2}{t})$$

Proof. We proof this result by induction.

Step 1: Initialization. For $t = 3$, we have by Lemma 7

$$\hat{\sigma}_3^2 = \frac{t-1}{t}\hat{\sigma}_2^2 + \frac{1}{3}(X_3 - \hat{\mu}_2)^2$$

As $X_1, X_2 \sim \mathcal{N}(\mu, \sigma^2)$ and are independent, we know that $\hat{\sigma}_2^2$ is independent of $\hat{\mu}_2$. We thus have $\hat{\mu}_2 \Big| \hat{\sigma}_2^2 \sim \mathcal{N}(\mu, \sigma^2/2)$.

Note that X_3 is independent from X_1 and X_2, so from σ_2^2 as well and we have $X_3 \Big| \hat{\sigma}_2^2 \sim \mathcal{N}(\mu, \sigma^2)$.

This means that $U = (X_3 - \hat{\mu}_2)\Big|\hat{\sigma}_2^2$ and $V = (\hat{\mu}_3 - \mu)\Big|\hat{\sigma}_2^2 = (\frac{1}{3}X_3 + \frac{2}{3}\hat{\mu}_2 - \mu)\Big|\hat{\sigma}_2^2$ are zero-mean gaussians.

Note also that $\hat{\mu}_2 \Big| \hat{\sigma}_2^2$ is independent of $X_3 \Big| \hat{\sigma}_2^2$, again because X_3 is independent of (X_1, X_2). From that we deduce that

$$\mathbb{E}\Big[UV\Big] = \mathbb{E}\Big[\Big(X_3 - \hat{\mu}_2\Big)\Big(\frac{1}{3}X_3 + \frac{2}{3}\hat{\mu}_2 - \mu\Big)\Big|\hat{\sigma}_2^2\Big]$$
$$= \mathbb{E}\Big[\Big((X_3 - \mu) - (\hat{\mu}_2 - \mu)\Big)\Big(\frac{1}{3}(X_3 - \mu) + \frac{2}{3}(\hat{\mu}_2 - \mu)\Big)\Big|\hat{\sigma}_2^2\Big]$$
$$= \frac{1}{3}\sigma^2 - \frac{2}{3}\frac{\sigma^2}{2} = 0.$$

As U and V are zero mean gaussians, the fact that their cross product is zero means that they are independent and that $V \Big| U = V \Big| \Big\{\hat{\sigma}_2^2, U\Big\} \sim V \Big| \Big\{\hat{\sigma}_2^2\Big\}$. In other words,

$$(\hat{\mu}_3 - \mu)\Big|\Big\{\hat{\sigma}_2^2, (X_3 - \hat{\mu}_2)^2\Big\} \sim (\hat{\mu}_3 - \mu)\Big|\Big\{\hat{\sigma}_2^2\Big\} \sim \mathcal{N}(\mu, \sigma^2/3),$$

because $(\hat{\mu}_3 - \mu)\Big|\sigma_2^2 \sim \mathcal{N}(0, \sigma^2/3)$ from the fact that $\hat{\mu}_2\Big|\hat{\sigma}_2^2 \sim \mathcal{N}(\mu, \sigma^2/2)$ and $X_3\Big|\hat{\sigma}_2^2 \sim \mathcal{N}(\mu, \sigma^2/2)$ and because they are independent.

This shows the property for $t = 3$.

Step 2: Induction Property. Assume that for a given t, we have

$$\hat{\mu}_t\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\} \sim \mathcal{N}(\mu, \frac{\sigma^2}{t})$$

Let us call $U = (X_{t+1} - \hat{\mu}_t)\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\}$ and $V = (\mu_{t+1} - \mu)\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\} = (\frac{t}{t+1}(\mu_t - \mu) + \frac{t}{t+1}(X_{t+1} - \mu))\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\}$.

Note that because of X_{t+1} is independent from $(X_1, \ldots, X_t)$, we have

$$X_{t+1}\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\} \sim \mathcal{N}(\mu, \sigma^2),$$

and also that $X_{t+1}\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\}$ is independent from $\hat{\mu}_t\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\}$. This combined with the induction assumption ensures that U and V are zero-mean gaussians, and also that

$$\begin{aligned}
\mathbb{E}\Big[UV\Big] &= \mathbb{E}\Big[\Big(X_{t+1} - \hat{\mu}_t\Big)\Big(\frac{1}{t+1}X_{t+1} + \frac{t}{t+1}\hat{\mu}_t - \mu\Big)\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\}\Big] \\
&= \mathbb{E}\Big[\Big((X_{t+1} - \mu) - (\hat{\mu}_t - \mu)\Big)\Big(\frac{1}{t+1}(X_{t+1} - \mu) + \frac{t}{t+1}(\hat{\mu}_t - \mu)\Big)\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\}\Big] \\
&= \frac{1}{t+1}\sigma^2 - \frac{t}{t+1}\frac{\sigma^2}{t} = 0
\end{aligned}$$

As U and V are zero mean gaussians, the fact that their cross product is zero means that they are independent and that $V\Big|U = V\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}, U\Big\} \sim V\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\}$. In other words,

$$(\hat{\mu}_{t+1} - \mu)\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t}\Big\} \sim (\hat{\mu}_{t+1} - \mu)\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\} \sim \mathcal{N}(\mu, \sigma^2/(t+1)),$$

because $(\hat{\mu}_{t+1} - \mu)\Big|\sigma_2^2 \sim \mathcal{N}(0, \sigma^2/(t+1))$ from the fact that $\hat{\mu}_t\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\} \sim \mathcal{N}(\mu, \sigma^2/t)$ and $X_{t+1}\Big|\Big\{\hat{\sigma}_2^2, \Big((X_{i+1} - \hat{\mu}_i)^2\Big)_{i \le t-1}\Big\} \sim \mathcal{N}(\mu, \sigma^2/2)$ and because they are independent.

This finishes the induction.

From Lemma 8 and the second part of Lemma 7, we have $\forall 2 \leq t \leq n$:

$$\hat{\mu}_t \Big| \left\{ \left(\hat{\sigma}_i^2 \right)_{i \leq t} \right\} \sim \mathcal{N}(\mu, \frac{\sigma^2}{t})$$

The algorithm we use is such that $T_{k,n}$ is a (noisy with respect to the samples on the other arms) function of the $\left(\hat{\sigma}_{k,t}^2\right)_{t \leq n}$ and that $T_{k,n}$ depends on the samples of arm k only trough those empirical variances $\left(\hat{\sigma}_{k,t}^2\right)_{t \leq n}$.

We thus have

$$\begin{aligned}
\mathbb{E}\Big((\hat{\mu}_{k,n} - \mu_k)^2 \mathbb{I}\{\xi\} \Big) &= \sum_{t=\underline{T_{k,n}}}^{n} \mathbb{E}\Big((\hat{\mu}_{k,n} - \mu_k)^2 \Big| T_{k,n} = t \Big) \mathbb{P}(T_{k,n} = t) \\
&= \sum_{t=\underline{T_{k,n}}}^{n} \sum_{(b_i^2)_{i \leq t}} \mathbb{E}\Big((\hat{m}_t - \mu_k)^2 \Big| \left\{ \left(s_i^2 \right)_{i \leq t} = (b_i^2)_{i \leq t} \right\} \Big) \\
&\times \mathbb{P}\Big[\left\{ \left(s_i^2 \right)_{i \leq t} = \left(b_i^2 \right)_{i \leq t} \right\} \Big| T_{k,n} = t \Big] \mathbb{P}(T_{k,n} = t) \\
&= \sum_{t=\underline{T_{k,n}}}^{n} \sum_{(b_i^2)_{i \leq t}} \frac{\sigma_k^2}{t} \mathbb{P}\Big[\left\{ \left(s_i^2 \right)_{i \leq t} = \left(b_i^2 \right)_{i \leq t} \right\} \Big| T_{k,n} = t \Big] \mathbb{P}(T_{k,n} = t) \\
&\leq \frac{\sigma_k^2}{\underline{T_{k,n}}},
\end{aligned}$$

where $\hat{m}_t = \frac{1}{t} \sum_{i=1}^{t} X_i$ where X_i is the ith sample collected from arm k.

□

We prove Theorem 3.

Proof (Theorem 3). Note that Lemma 2 is only based on the assumption that samples are generated by a sub-Gaussian distribution. Here we strengthen that assumption and require all the distributions to be Gaussian with parameters μ_k and σ_k^2. We recall that the loss of any arm k is decomposed in two parts as follows:

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

From Lemma 3 and the bound in Equation 58, we have

$$
\begin{aligned}
&\mathbb{E}[(\hat{\mu}_{k,n}-\mu)^2\mathbb{I}\{\xi\}] \leq \frac{\sigma_k^2}{\underline{T}_{k,n}} \\
&\leq \frac{\Sigma}{n} + \frac{4K}{n^2} + \frac{B}{n^{3/2}} + \frac{C}{n^{7/4}} \\
&\leq \frac{\Sigma}{n} + \frac{4K}{n^2} + \frac{7056}{n^{3/2}}[3528\Big(c_1(c_2+2)^2+1\Big)K^{3/2}\log(n)^2 \\
&+ 1280K^2\Big(c_1(c_2+2)^2\log(n)^2 + \frac{(\sqrt{c_2}+1)\bar{\mu}^2}{n^2}\Big)n^{-7/4} \\
&\leq \frac{\Sigma}{n} + \frac{4K}{n^2} + \frac{7056}{n^{3/2}}\Big(c_1(c_2+2)^2+1\Big)K^{3/2}\log(n)^2 \\
&+ 2560K^2\sqrt{\Sigma(c_2+1)}(\bar{\mu}+\bar{\mu}^2) + c_1(c_2+2)^2\log(n)^2 n^{-7/4}. \qquad (61)
\end{aligned}
$$

where we use the bounds on B and C in Appendix B.1. Note that for Gaussian distributions $\mathbb{P}(|X-\mu| > t) \leq \exp(-t^2/2\sigma^2)$. From Lemma 6 and by taking $c_1 = 2\Sigma$ and $c_2 = 1$, we obtain also with $\delta = n^{-5/2}$

$$
\mathbb{E}\big[(\hat{\mu}_{k,n}-\mu)^2\mathbb{I}\Big\{\xi^C\Big\}\big] \leq \frac{8K}{n^{3/2}}\log(n). \qquad (62)
$$

Finally, combining Equations 61 and 62, and recalling the definition of regret, we have

$$
\begin{aligned}
R_n \leq\ & \frac{4K}{n^2} + \frac{7060}{n^{3/2}}\Big(c_1(c_2+2)^2+1\Big)K^{3/2}\log(n)^2 \\
&+ 2560K^2\sqrt{\Sigma(c_2+1)}(c_1(c_2+2)^2\log(n)^2 n^{-7/4}\ . \qquad (63)
\end{aligned}
$$