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] 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 variance 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 distribution.

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 step 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 advertisement 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 level of 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 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

J. Kivinen et al. (Eds.): ALT 2011, LNAI 6925, pp. 189–203, 2011.

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 option (or arm) k_t and as a consequence receives a random sample from ν_{k_t} (independent of the past samples). Given a budget of n samples, the goal is to define an allocation strategy over arms so as to estimate their expected values uniformly well (using a squared loss to evaluate the accuracy). 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 precisely, proportionally to $\lambda_k = \sigma_k^2 / \sum_j \sigma_j^2$. However, since the distributions are initially unknown, the learner should implement an active allocation strategy which adapts its behavior as samples are collected. The performance of this strategy is measured by its regret (Eq. 4), defined as 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.[1] 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 may have a poor performance whenever an arm has a very small variance compared to the others (e.g., when users involved in the poll have very similar opinions about some products and very different on some 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.

In this paper, in order to further investigate this issue, we introduce two novel algorithms based on upper-confidence-bounds (UCB) on the variance. The algorithms sample the arms proportionally to an upper-bound on their variance computed from the empirical variances and a confidence interval derived from Chernoff-Hoeffding's (first algorithm) and Bernstein's (second algorithm) inequalities. The main advantage of this class of algorithms is that the possibility to use standard tools and arguments for UCB-like algorithms makes their analysis simple, thus making the study of the dependency on $\lambda_{\min}$ easier. The main contributions and findings of this paper are as follows:

– The first algorithm, called CH-AS, is based on Chernoff-Hoeffding's bound and its regret is $\tilde{O}(n^{-3/2})$ with an inverse dependency on $\lambda_{\min}$, similar to GAFS-MAX. The main differences are: the bound for CH-AS holds for any n

[1] 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.

(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 it 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 again in the bound. 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 from $\lambda_{\min}$ for the case of Gaussian arms.
- 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. Furthermore, we further investigate the impact of the distribution on the regret by reporting numerical results in case of Rademacher distributions showing that B-AS performance worsens with $\lambda_{\min}^{-1}$. This leads to the conjecture 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 in the standard K-armed stochastic bandit setting, where each arm $k = 1, \dots, K$ is characterized by a distribution[2] 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

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

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

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

[2] Although the formulation of the problem holds for any distribution, in the following we will consider the case of bounded and sub-Gaussian distributions in order to derive meaningful bounds.

If the variance of the arms were known in advance, one could design an optimal static allocation (i.e., independent from the observed samples) by pulling the arms proportionally to their variances. If an arm k is pulled a fixed number of times $T^*_{k,n}$, its loss is [3]

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

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 $T^*_{k,n} = \frac{\sigma_k^2 n}{\sum_{i=1}^K \sigma_i^2}$ times (up to rounding effects), 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). In fact, in both settings good strategies should play each arm a linear function of n, in contrast with the standard stochastic bandit setting, where the sub-optimal arms should be played logarithmically in n.

3 Allocation Strategy Based on Chernoff-Hoeffding UCB

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

[3] This equality does not hold when the number of pulls is random, e.g., in adaptive algorithms, where the strategy depends on the random observed samples.

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

Fig. 1. The pseudo-code of the CH-AS algorithm, with $\hat{\sigma}^2_{q,t}$ computed as in Eq. 5

3.1 The CH-AS Algorithm

The CH-AS algorithm $\mathcal{A}_{CH}$ in Fig. 1 takes a confidence parameter δ as input and after n pulls returns an empirical mean $\hat{\mu}_{q,n}$ for each arm q. At each time step t, i.e., after having pulled arm k_t, the algorithm computes the empirical mean $\hat{\mu}_{q,t}$ and variance $\hat{\sigma}^2_{q,t}$ of each arm q as[4]

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

where $X_{q,i}$ is the i-th sample of ν_q 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 $2K$), 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:

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

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

$$-\frac{5}{\Sigma^2 \lambda_{\min}^{3/2}}\sqrt{n\log(1/\delta)} - \frac{K}{\Sigma} \le T_{k,n} - T^*_{k,n} \le \frac{5(K-1)}{\Sigma^2 \lambda_{\min}^{3/2}}\sqrt{n\log(1/\delta)} + \frac{K^2}{\Sigma} \,. \tag{6}$$

[4] Notice that this is a biased estimator of the variance even if the $T_{q,t}$ were not random.

Proof. Let $\xi_{K,n}(\delta)$ be the event

$$\xi_{K,n}(\delta) = \bigcap_{1\leq k\leq K,\ 1\leq t\leq n} \left\{ \left| \left(\frac{1}{t}\sum_{i=1}^{t} X_{k,i}^2 - \left(\frac{1}{t}\sum_{i=1}^{t} X_{k,i}\right)^2 \right) - \sigma_k^2 \right| \leq 5\sqrt{\frac{\log(1/\delta)}{2t}} \right\}. \quad (7)$$

From Hoeffding's inequality it follows that $\Pr(\xi_{K,n}(\delta)) \geq 1 - 4nK\delta$. We divide the proof of this lemma into the following three steps.

Step 1. Mechanism of the algorithm. On event $\xi_{K,n}(\delta)$, for all $t \leq n$ and q

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

and the following upper and lower bounds for $B_{q,t+1}$ hold

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

Let $t+1 > 2K$ be the time at which a given arm k is pulled for the last time, i.e., $T_{k,t} = T_{k,n} - 1$ and $T_{k,(t+1)} = T_{k,n}$. Note that as $n \geq 4K$, there is at least one arm k that is pulled after the initialization phase. Since $\mathcal{A}_{CH}$ chooses to pull arm k at time $t+1$, for any arm p, we have $B_{p,t+1} \leq B_{k,t+1}$. From Eq. 8 and the fact that $T_{k,t} = T_{k,n} - 1$, we obtain

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

Using Eq. 8 and the fact that $T_{p,t} \leq T_{p,n}$, we derive a lower-bound for $B_{p,t+1}$ as

$$B_{p,t+1} \geq \frac{\sigma_p^2}{T_{p,t}} \geq \frac{\sigma_p^2}{T_{p,n}}\ . \quad (10)$$

Combining the condition $B_{p,t+1} \leq B_{k,t+1}$ with Eqs. 9 and 10, we obtain

$$\frac{\sigma_p^2}{T_{p,n}} \leq \frac{1}{T_{k,n}-1}\left(\sigma_k^2 + 10\sqrt{\frac{\log(1/\delta)}{2(T_{k,n}-1)}}\right). \quad (11)$$

Note that at this point there is no dependency on t, and thus, the probability that Eq. 11 holds for any p and for any k such that $T_{k,n} > 2$ (i.e. arm k is pulled at least once after the initialization phase), is at least $1 - 4nK\delta$ (probability of the event $\xi_{K,n}(\delta)$).

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

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

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

$$T_{p,n} \geq \frac{\sigma_p^2}{\frac{\Sigma}{n} + \frac{20\sqrt{\log(1/\delta)}}{(n\lambda_{\min})^{3/2}} + \frac{4K\Sigma}{n^2}} \geq T^*_{p,n} - \frac{5\sqrt{n\log(1/\delta)}}{\Sigma^2 \lambda_{\min}^{3/2}} - \frac{K}{\Sigma}, \tag{12}$$

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. 12 and the fact that $\sum_k T_{k,n} = n$, we obtain the upper bound

$$T_{p,n} = n - \sum_{k \neq p} T_{k,n} \leq T^*_{p,n} + \frac{5(K-1)}{\Sigma^2 \lambda_{\min}^{3/2}} \sqrt{n\log(1/\delta)} + \frac{K^2}{\Sigma}\,. \tag{13}$$

□

We now show how this bound translates into a regret bound.

Theorem 1. *Assume the distributions $\{\nu_k\}_{k=1}^K$ to be bounded in $[0,1]$ and $n \geq 4K$. The regret of $\mathcal{A}_{CH}$, for parameter $\delta = n^{-5/2}$, is bounded as*

$$R_n(\mathcal{A}_{CH}) \leq \frac{70K\sqrt{\log n}}{n^{3/2}\ \Sigma\ \lambda_{\min}^{5/2}} + O\Big(\frac{\log n}{n^2}\Big)\,. \tag{14}$$

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 often an arm is pulled, the smaller its estimation error becomes. However, this is not true in general because $T_{k,n}$ is a random variable that depends on the actual received samples, and thus, $L_{k,n} = \mathbb{E}_{\nu_k}\big[(\mu_k - \hat{\mu}_{k,n})^2\big]$ does not satisfy Eq. 3. Nevertheless, for any arm k, the number of pulls $T_{k,n}$ is a stopping time w.r.t. the filtration induced by the samples received for arm k. Hence, by applying the result of Lemma 10 in Antos et al. (2010) (a form of Wald's equality), one derive

$$\mathbb{E}\big[(\mu_k - \hat{\mu}_{k,n})^2 \mathbb{I}\left\{\xi_{K,n}(\delta)\right\}\big] \leq \frac{1}{\underline{T}^2_{k,n}} \mathbb{E}\Big[\Big(\sum_{t=1}^{T_{k,n}} (\mu_k - X_{k,t})\Big)^2\Big] = \frac{\sigma_k^2 \mathbb{E}(T_{k,n})}{\underline{T}^2_{k,n}}\,, \tag{15}$$

where $\underline{T}_{k,n}$ is a lower-bound for $T_{k,n}$ on $\xi_{K,n}(\delta)$. From this bound, one can use Lemma 1, which provides both upper and lower-bounds for $T_{k,n}$ on the event $\xi_{K,n}(\delta)$ to deduce that $\mathbb{E}\big[(\mu_k - \hat{\mu}_{k,n})^2 \mathbb{I}\left\{\xi_{K,n}(\delta)\right\}\big] = \frac{\sigma_k^2}{T^*_{k,n}} + O\big(n^{-3/2}\sqrt{\log(1/\delta)}\big)$

and $\mathbb{E}\left[(\mu_k - \hat{\mu}_{k,n})^2 \mathbb{I}\left\{\xi_{K,n}(\delta)\right\}^c\right] \leq 1 \times \mathbb{P}(\xi_{K,n}(\delta)^c) \leq 4nK\delta$ (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 optimal allocation strategy, which assumes that the variances 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 for GAFS-MAX algorithm (Antos et al., 2010).

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 1 that the inverse dependency on $\lambda_{\min}$ appears in the bound on the number of pulls and then it is propagated to the regret bound. We now show with a simple example that this dependency is not an artifact of the analysis and it 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 an unbounded regret but 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 upper-confidence term forces the algorithm to pull the arm with zero variance at least $\Omega(n^{2/3})$ 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 faster than $1/n$, but with a slower rate than in Thm. 1. Merging these two results, we deduce that the regret is in fact $R_n \leq \min\left\{\lambda_{\min}^{-5/2}\tilde{O}(n^{-3/2}), \tilde{O}(n^{-4/3})\right\}$. Note that when $\lambda_{\min} = 0$ the regret of GAFS-MAX is in $\tilde{O}(n^{-3/2})$[5], and GAFS-MAX thus outperforms CH-AS in this case. We further study the behavior of CH-AS in Sec. 5.1.

The reason for the poor performance in Lemma 1 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. 18 and 19).

[5] See the end of Section 4 in Antos et al. (2010).

Input: parameters c_1, c_2, δ
Let $b = 4\sqrt{c_1 \log(c_2/\delta)}\sqrt{\log(2/\delta)} + 2\sqrt{5c_1}n^{-1/2}$
Initialize: Pull each arm twice
for $t = 2K+1, \ldots, n$ **do**
  Compute $B_{q,t} = \frac{1}{T_{q,t-1}}\Big(\hat{\sigma}^2_{q,t-1} + 2b\hat{\sigma}_{q,t-1}\sqrt{\frac{1}{T_{q,t-1}}} + b^2\frac{1}{T_{q,t-1}}\Big)$ 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: $\hat{\mu}_{q,t}$ for each arm $1 \le q \le K$

Fig. 2. The pseudo-code of the B-AS algorithm, with $\hat{\sigma}_{k,t}$ computed as in Eq. 16

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, with an improved bound on $|T_{k,n} - T^*_{k,n}|$ without the inverse dependency on $\lambda_{\min}$ (compare the bounds in Eqs. 18 and 19 to the one for CH-AS in Eq. 6). However this result itself is not sufficient to derive a better regret bound than CH-AS. This finding 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 the number of pulls translates into a regret bound depends on the specific distributions of the arms. In fact, when the sample distributions are Gaussian, we can exploit the property that the empirical mean $\hat{\mu}_{k,t}$ conditioned on $T_{k,t}$ is independent of the empirical variances $(\hat{\sigma}_{k,s})_{s\le t}$ and further deduce that the regret of B-AS no longer depends on $\lambda^{-1}_{\min}$. The numerical simulations in Sec. 5 further illustrate this theoretical finding.

4.1 The B-AS Algorithm

The B-AS algorithm (Fig. 2), $\mathcal{A}_B$, is based on a high-probability bounds (empirical Bernstein's inequality) on the variance of each arm (Maurer and Pontil, 2009, Audibert et al., 2009). B-AS requires three parameters as input (see also Remark 4 in Sec. 4.2 on how to reduce them 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

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

where $b = 4\sqrt{c_1 \log(c_2/\delta)}\sqrt{\log(2/\delta)} + 2\sqrt{5c_1}n^{-1/2}$ and[6]

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

$$\hat{\sigma}_{k,t}^2 = \frac{1}{T_{k,t}-1}\sum_{i=1}^{T_{k,t}}(X_{k,i}-\hat{\mu}_{k,t})^2, \qquad \text{with} \qquad \hat{\mu}_{k,t} = \frac{1}{T_{k,t}}\sum_{i=1}^{T_{k,t}} X_{k,i}\ . \tag{16}$$

4.2 Regret Bound and Discussion

Instead of bounded distributions, we consider the more general assumption of sub-Gaussian distributions.

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

$$\mathbb{P}_{X\sim\nu_k}(|X-\mu_k| \ge \epsilon) \le c_2 \exp(-\epsilon^2/c_1)\ . \tag{17}$$

We first bound 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 $T_{p,\min} \le T_{p,n} \le T_{p,\max}$ for any arm $1 \le p \le K$ and any $n \ge \frac{16}{9}c(\delta)^{-2}$, where*

$$T_{p,\min} = T^*_{p,n} - K\lambda_p\left[1 + \frac{16a\sqrt{\log(2/\delta)}}{\Sigma}\left(\sqrt{\Sigma} + \frac{2a\sqrt{\log(2/\delta)}}{c(\delta)}\right)\sqrt{n} + 128Ka^2\frac{\log(2/\delta)}{\Sigma\sqrt{c(\delta)}}\,n^{1/4}\right], \tag{18}$$

and

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

where $c(\delta) = \frac{2a\sqrt{\log(2/\delta)}}{\sqrt{K}(\sqrt{\Sigma}+4a\sqrt{\log(2/\delta)})}$ and $a = 2\sqrt{c_1\log(c_2/\delta)} + \sqrt{\frac{5c_1}{\log(2/\delta)}}n^{-1/2}$.

Remark. Unlike the bounds for CH-AS in Lemma 1, B-AS allocates the arms such that the difference between $T_{p,n}$ and $T^*_{p,n}$ grows at most as $\tilde{O}(\sqrt{n})$ without dependency on $\lambda_{\min}^{-1}$. 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 $O(n^{2/3})$ whenever $\lambda_{\min}$ is small. We further notice that the lower bound in Eq. 18 is of order $\lambda_p\tilde{O}(\sqrt{n})$, 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. *Under Assumption 1, for any $n \ge 4K$, the regret of $\mathcal{A}_B$ run with parameters c_1, c_2, and $\delta = (c_2+2)n^{-5/2}$ is bounded as*

$$R_n(\mathcal{A}_B) \le \Big[\frac{CK^5c_1^2}{\lambda_{\min}^2}\log(n)^2\Big(\Sigma + 200\log(n)\Big)\Big(1 + \frac{1}{\Sigma^3}\Big) + 2c_1(c_2+2)K\Big]n^{-3/2}\ ,$$

where C is a constant (a loose numerical value for C is 30000).

Similar to Thm. 1, the bound on the number of pulls translates into a regret bound through Eq. 15. As it can be noticed, in order to remove the dependency on $\lambda_{\min}$, a symmetric bound on $|T_{p,n} - T^*_{p,n}| \leq \lambda_p \tilde{O}(\sqrt{n})$ would be needed. While the lower bound in Eq. 18 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. In the next section, we show that an improved regret bound can be achieved in the special case of Gaussian distributions.

4.3 Regret for Gaussian Distributions

In the case of Gaussian distributions, the loss bound in Eq. 15 can be improved as in the following lemma (the full proof is reported in Carpentier et al. 2011).

Lemma 3. *Assume that distributions $\{\nu_k\}_{k=1}^K$ are Gaussian. Then for any k*

$$\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big] \leq \frac{\sigma_k^2}{T_{k,min}} + \sigma_k^2 \delta' \,, \tag{20}$$

where $T_{k,n} \geq T_{k,\min}$ is the lower-bound in Lemma 2 which holds with probability at least $1 - \delta'$ (where $\delta' = 2nK\delta$).

Proof (Sketch). We first write the loss for any arm k as

$$\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2\big] = \sum_{t=2}^{n} \mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2 | T_{k,n} = t\big] \mathbb{P}(T_{k,n} = t). \tag{21}$$

We notice that $T_{k,n}$ is a random stopping time which depends on the sequence of empirical variances for arm k and the empirical variances of all the other arms. The event $\{T_{k,n} \geq t\}$ depends on the filtration $\mathcal{F}_{k,t}$ (generated by the sequence of empirical variances of the rewards of arm k) and on the "environment of arm k" $\mathcal{E}_{-k}$ (defined by all the rewards samples of other arms). We recall that for a Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k^2)$, the empirical mean $\hat{\mu}_{k,n}$ built on a fixed number t of independent samples is distributed as a normal distribution $\mathcal{N}(\mu_k, \sigma_k^2/t)$ and it is independent from the empirical variance $\hat{\sigma}^2_{k,n}$. According to Carpentier et al. (2011), this property can be extended to the conditional random variable $\hat{\mu}_{k,n} | \mathcal{F}_{k,n}, \mathcal{E}_{-k}$ which is still distributed as $\mathcal{N}(\mu_k, \sigma_k^2/t)$. Using this property in (21) we have

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

Using the lower-bound in Lemma 2 the statement follows. □

Remark 1. We notice that the loss bound in Eq. 20 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 a 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 $\{\hat{\sigma}_k^2\}_k$ (as in CH-AS and B-AS), the conditional expectation $\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2 | T_{k,n} = t\big]$ no longer equals

σ_k^2/t. However, for Gaussian distributions we recover the property $\mathbb{E}\big[(\hat{\mu}_{k,n} - \mu_k)^2 | T_{k,n} = t\big] = \sigma_k^2/t$, which allows us to deduce the result 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 $c_2 = 1$.[7]

Theorem 3. *Assume that $\{\nu_k\}_{k=1}^K$ are Gaussian and that an upper-bound $\overline{\Sigma} \geq \Sigma$. B-AS with parameters $c_1 = 2\overline{\Sigma}$, $c_2 = 1$, and $\delta = n^{-5/2}$ has a regret*

$$R_n(\mathcal{A}_B) \leq C\overline{\Sigma}K^{3/2}\big(\log(2n)\big)^2 n^{-3/2} + O\Big(n^{-7/4}(\log n)^2\Big), \tag{22}$$

where C is a constant (a loose numerical value for C is 19200).

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, 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 on a two-armed problem with Gaussian distributions $\nu_1 = \mathcal{N}(0, \sigma_1^2 = 4)$ and $\nu_2 = \mathcal{N}(0, \sigma_2^2 = 1)$ (note that $\lambda_{\min}$=1/5). Fig. 3-*(left)* shows the rescaled regret, $n^{3/2}R_n$, for the three algorithms averaged over 50,000 runs. The results indicate that while the rescaled regret is almost constant 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. 3 (Eq. 22). The initial increase in the CH-AS curve is also consistent with the bound of Thm. 1 (Eq. 14). As discussed in Remark 3 of Sec. 3.2, the regret bound for CH-AS is of the form $R_n \leq \min\big\{\lambda_{\min}^{-5/2}\tilde{O}(n^{-3/2}), \tilde{O}(n^{-4/3})\big\}$, and thus, the algorithm behaves as $\tilde{O}(n^{-4/3})$ and $\lambda_{\min}^{-5/2}\tilde{O}(n^{-3/2})$ for small and large (relative to $\lambda_{\min}^{-1}$) values of n, respectively. It is important to note that the behavior of CH-AS is independent of the arms' distributions and is intrinsic in the allocation mechanism, as shown in Lemma 1. Finally, the behavior of GAFS-MAX indicates that although its analysis shows an inverse dependency on $\lambda_{\min}$ and yields a regret bounds similar to CH-AS,

[7] For a single Gaussian distribution $c_1 = 2\sigma^2$. Here we use $c_1 = 2\Sigma$ in order for the assumption to be satisfied for all K distributions simultaneously.

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. This would remove the inverse dependency on $\lambda_{\min}$ and provide a regret bound similar to B-AS in the case of Gaussian distributions.

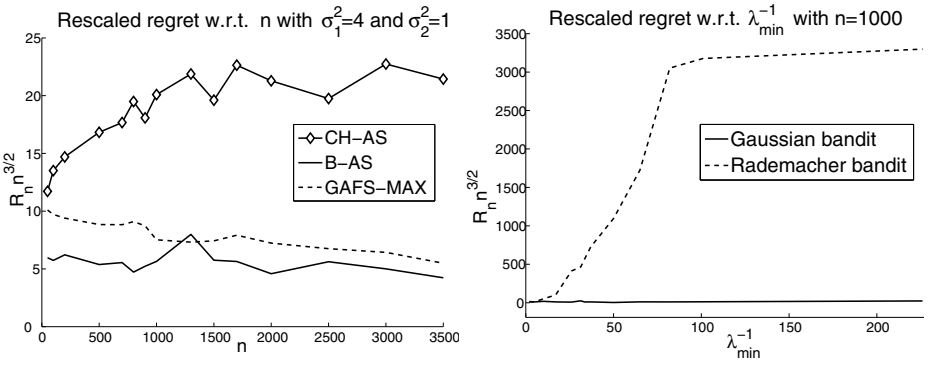

Fig. 3. *(left)* The rescaled regret of CH-AS, B-AS, and GAFS-MAX algorithms on a two-armed problem, where the distributions of the arms are Gaussian. *(right)* The rescaled regret of B-AS for two bandit problems, one with two Gaussian arms and one with a Gaussian and a Rademacher arms.

5.2 B-AS with Non-gaussian Arms

In Sec. 4.3, we showed that when the arms have Gaussian distributions, the regret bound of the B-AS algorithm does not depend on $\lambda_{\min}$ anymore. We also discussed 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 directly impacts 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 $\hat{\sigma}^2_{k,n}$ conditioned on $T_{k,n}$. 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,[8] the loss $(\hat{\mu}_{k,t} - \mu_k)^2$ is equal to $\hat{\mu}^2_{k,t}$ and we have $\hat{\sigma}^2_{k,t} = \frac{1}{T_{k,t}} \sum_{i=1}^{T_{k,t}} X^2_{k,i} - \hat{\mu}^2_{k,t} = 1 - \hat{\mu}^2_{k,t}$, as a result, the larger $\hat{\sigma}^2_{k,t}$, the smaller $\hat{\mu}^2_{k,t}$. 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 $\hat{\sigma}^2_{k,t}$,

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

the more often arm k is pulled. In case of Rademacher distributions, this means that an arm is pulled more than its optimal allocation exactly when its mean is accurately estimated (the loss is small). This may result in a poorer estimation of the arm, and thus, negatively affect the regret of the algorithm.

In the experiments of this section, we use B-AS in two different bandit problems: one with two Gaussian arms $\nu_1 = \mathcal{N}(0, \sigma_1^2)$ (with $\sigma_1 \geq 1$) and $\nu_2 = \mathcal{N}(0, 1)$, and one with a Gaussian $\nu_1 = \mathcal{N}(0, \sigma_1^2)$ 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 the uniform estimation of the mean value of K independent distributions under a given sampling budget. We introduced a novel class of algorithms based on upper-confidence-bounds on the (unknown) variances of the arms, and analyzed two algorithms: CH-AS and B-AS. For CH-AS we derived a regret bound similar to Antos et al. (2010), scaling as $\tilde{O}(n^{-3/2})$ and with a dependence on $\lambda_{\min}^{-1}$. We then introduced a more refined algorithm, B-AS, using a tighter upper bounds on the variance, and reported a refined regret bound in the case of Gaussian distributions. Finally we gave arguments (including numerical simulations) supporting the idea that the full shape of the distributions (and not not only their variance) has a relevant impact on the performance of the allocation strategies.

This work opens a number of questions.

- *Distribution dependency.* An 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 sample 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 $\lambda_{\min}^{-1}$).

Acknowledgements. We thank András Antos for many comments that helped us to greatly improve the quality of the paper. This work was supported by French National Research Agency (ANR-08-COSI-004 project EXPLO-RA) and by the European Community's Seventh Framework Programme (project COMPLACS, grant agreement n°231495).

References

Antos, A., Grover, V., Szepesvári, C.: Active learning in heteroscedastic noise. Theoretical Computer Science 411, 2712–2728 (2010)

Audibert, J.-Y., Munos, R., Szepesvari, C.: Exploration-exploitation trade-off using variance estimates in multi-armed bandits. Theoretical Computer Science 410, 1876–1902 (2009)

Audibert, J.-Y., Bubeck, S., Munos, R.: Best arm identification in multi-armed bandits. In: Proceedings of the Twenty-Third Annual Conference on Learning Theory (COLT 2010), pp. 41–53 (2010)

Bubeck, S., Munos, R., Stoltz, G.: Pure exploration in finitely-armed and continuous-armed bandits. Theoretical Computer Science 412, 1832–1852 (2011) ISSN 0304-3975

Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R., Auer, P.: Upper-confidence-bound algorithms for active learning in multi-armed bandits. Technical Report inria-0059413, INRIA (2011)

Castro, R., Willett, R., Nowak, R.: Faster rates in regression via active learning. In: Proceedings of Neural Information Processing Systems (NIPS), pp. 179–186 (2005)

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

Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. J. Artif. Int. Res. 4, 129–145 (1996) ISSN 1076-9757

Étoré, P., Jourdain, B.: Adaptive optimal allocation in stratified sampling methods. Methodology and Computing in Applied Probability 12, 335–360 (2010)

Fedorov, V.: Theory of Optimal Experiments. Academic Press, London (1972)

Maurer, A., Pontil, M.: Empirical bernstein bounds and sample-variance penalization. In: Proceedings of the Twenty-Second Annual Conference on Learning Theory,7 pp. 115–124 (2009)