# Bridging the gap between regret minimization and best arm identification, with application to A/B tests

**Rémy Degenne**
LPSM, Université Paris Diderot,
CMLA, ENS Paris Saclay

**Thomas Nedelec**
Criteo AI Lab,
CMLA, ENS Paris Saclay

**Clément Calauzènes**
Criteo AI Lab

**Vianney Perchet**
CMLA, ENS Paris Saclay,
Criteo AI Lab

## Abstract

State of the art online learning procedures focus either on selecting the best alternative ("best arm identification") or on minimizing the cost (the "regret"). We merge these two objectives by providing the theoretical analysis of cost minimizing algorithms that are also $\delta$-PAC (with a proven guaranteed bound on the decision time), hence fulfilling at the same time regret minimization and best arm identification. This analysis sheds light on the common observation that ill-callibrated UCB-algorithms minimize regret while still identifying quickly the best arm.

We also extend these results to the non-iid case faced by many practitioners. This provides a technique to make cost versus decision time compromise when doing adaptive tests with applications ranging from website A/B testing to clinical trials.

## Introduction

With the growing use of personalization and machine learning techniques on user-facing systems, randomized experiments – or A/B tests – have become a standard tool to evaluate the performances of different versions of the systems. Two of the main drawbacks of such experiment are its possible length and its cost, as it commonly takes few weeks or even months to ensure a statistically founded decision. Thus, lots of attention have been paid to the complexity of the underlying statistical tests [18, 13, 24, 10, 23, 24].

These approaches have indubitable practical interest, but are often limited to quite simple A/B frameworks because of their restrictive assumptions. The first restrictive one is the assumption that a good procedure should minimize the time needed to take a statistically significant decision [13, 24]. In many situations, the main objective of practitioners is to minimize the overall cost of A/B testing without preventing him to take the right decision given a certain time budget.

Another traditional assumption in the online learning literature is that outcomes arriving over time are independent [13, 24]. Numerous common scenarii do not satisfy this and practitioners cannot benefit from the statistically efficient methods available in the iid setting. For instance, when an online retailer wants to A/B test two versions of its mobile application, it may not be able to consider the purchases as independents when customers are buying recurrently (or clicking repeatedly in the CTR optimization problem).

To address the first limitation, we exhibit a well-tuned variant of the UCB algorithm [2] that is both able to take a $\delta$-PAC decision in finite time and reaches a low regret. This algorithm can be taken as a tool for practitioners to interpolate between the tasks of *best arm identification* and *regret minimization*. Even if this objective has been briefly mentioned and/or observed empirically [9], we provide the first theoretical analysis of such algorithm.

To handle the second limitation, we extend the ideas developed in the iid case to a more complex setting that can handle units arriving through time and delivering rewards continuously during the test. We provide sample efficient statistical decision rules and guarantees to take decisions in finite time in this setting, highlighting the clear trade-off between *regret minimization* and *best arm identification*, even on more complex settings.

### Regret vs Best-Arm Identification in iid setting

*Framework.* We consider the classical multi-armed bandit problem with $K \geq 2$ arms or "population". At each time step $t$, the agent chooses an arm $i \in [K] := \{1, \ldots, K\}$ and observes a reward drawn from an unknown distribution $r_t^i$ with expectation $r^i$. We assume that each $r_t^i$ are $\sigma^2$-subGaussian, where the variance (proxy) $\sigma^2$ is known. We denote by $\pi_n \in [K]$ the sequence of random variable indicating which arm to pull at time $n \in \mathbb{N}$.

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

*Objective.* We consider both natural objectives of bandit problems. The first one corresponds to *regret minimization*. It consists in minimizing the cumulative regret

$$R(T) = T\max\left\{r^i\,;i\in[K]\right\} - \mathbb{E}\sum_{t=1}^{T} r_t^{\pi_t}$$

In *A/B testing*, when $K = 2$, minimizing the regret is the same as minimizing the cost of testing a new technology or the impact of a clinical trial on patients.

The second objective, matching the problem of *best arm identification* with *fixed confidence*, is to design an algorithm for a given confidence level $\delta$, that minimizes the worst-case number of sample $T_\delta$ needed for the algorithm to finish and to return the optimal arm with probability $1-\delta$. Using an algorithm for best arm identification in an A/B test gives a guarantee on the amount time necessary before being able to to take a statistically significant decision.

Intuitively, an algorithm that is optimal for regret minimization is sub-optimal for best arm identification because its exploration is too slow. The opposite is also true since the exploration of optimal best arm identification algorithms is too aggressive for regret minimization.

We aim at studying a family of algorithms that interpolate between these two objectives. Informally, with $\delta\in[0,1]$, our objective is to design algorithms for which with probability $1-\delta$, for all bandit problems $P$ in some class, the worse arm is discarded after $T_\delta$ stages and we have both

$$T_\delta \le f(P,\delta) \quad \text{and} \quad R(T_\delta)\le g(P,\delta)\,.$$

The values $f(P,\delta)$ and $g(P,\delta)$ characterize the performances of the algorithm and should be as small as possible.

### Literature review

*Regret minimization.* This objective has been extensively studied in the bandit literature since the seminal paper of [22]. We mention two particular classes of well-known algorithms that we will use throughout the paper.

- The *Upper Confidence Bound* (UCB) algorithm introduced in [12, 2] decides which arms to consider depending on the respective empirical means and an error term depending on the number of pulls of each arm. Its regret, in the case of two arms with Gaussian rewards, is equal to $2\log(T)/\Delta$ where $\Delta = |r^{\mathcal{A}} - r^{\mathcal{B}}|$ is the gap between the mean of the two arms [2].
- The other class of algorithms we consider is known as *Explore Then Commit* (ETC)[20, 7]. They are first considering stages of pure exploration before exploiting the arm with the highest empirical mean. The algorithm consists in tuning the switching times between the stages. Its regret in the case of two arms with Gaussian rewards is of order $4\log(T)/\Delta$.

We recall that ETC is necessarily sub-optimal for *regret minimization*, as in the case of Gaussian rewards there exists a sub-optimal additional and multiplicative factor 2 [7].

*Best arm identification.* The problem of best arm identification [15], can be cast in two main settings depending on the constraint imposed on the system:

- *fixed budget* [1, 4] where a total number of samples $T\in\mathbb{N}$ is given and the goal is to minimize the error probability at time T;
- *fixed confidence* [15, 5] where the goal is to minimize the total number of stages used to return the best arm with probability $1-\delta$.

In the fixed confidence setting there are two main ways to evaluate the sample complexity of the algorithm : *the average sample complexity* studied in [15, 5, 13, 14] where the goal is to minimize the expected time of decision and *the worst case sample complexity* studied in [6, 11, 9] where the objective is to have a quantity $T_\delta\in\mathbb{N}$ as low as possible such that with probability $1-\delta$, the algorithm makes no mistake and the time of decision $\tau_d$ is below $T_\delta$. In the case of two arms, the optimal sampling strategy is to sample each arm uniformly and stop with a criterion similar to the one used in ETC [13].

*A/B testing* Most of the statistical literature on A/B testing [13, 24] has focused on the objective of minimizing the time necessary to take a statistical sufficient decision and to the best of our knowledge, there exists no work theoretically interpolating between the objectives of best arm identification and regret minimization.

## 1 Simultaneous Best-arm Identification and Regret Minimization

In this section, we construct a family of algorithms that minimizes regret while being $\delta$-PAC (with a proven guaranteed bound on the decision time), hence fulfilling at the same time the regret minimization and best-arm identification. For the sake of clarity, we are going to assume that there are only $k = 2$ populations, as in A/B testing, denoted by $\mathcal{A}$ and $\mathcal{B}$. The results for the general case $K > 2$ can actually be deduced almost immediately from those when $K = 2$.

Let us first recall that the algorithm with lowest decision time[14], given the variance of arms are identical, is ETC, which pulls both arms uniformly and, after pulling each arm $n$ times, returns the arm (e.g. $\mathcal{A}$) with highest empirical average $\hat{r}_n^{\mathcal{A}}$, if $\hat{r}_n^{\mathcal{A}} - \hat{r}_n^{\mathcal{B}} \ge \sqrt{\frac{4\sigma^2}{n}\log\left(\frac{\log^2(n)}{\delta}\right)}$ .

In the statement of our theorems, the usual Landau notation $o_\delta(1)$ stands for any function whose limit is, as $\delta$ goes to 0, equal to 0. Similarly, we will use $\widetilde{\delta} \le 23\delta\log(\frac{1}{\delta})$ instead of

$\delta$ for the sake of clarity. Exact values of $\widetilde{\delta}$, precise statements and proofs are mostly delayed to Appendix B.

The performances of ETC are now well understood.

**Proposition 1** ([19]). *With probability greater than $1-\widetilde{\delta}$, ETC returns the best arm at a stage $\tau_d \leq T_\delta$ where*

$$T_\delta \leq \frac{32\sigma^2}{\Delta^2} \log\left(\frac{1}{\delta}\right)\left(1+o_\delta(1)\right) .$$

*So the regret of ETC at the time of decision verifies*

$$R(\tau_d) \leq \frac{16\sigma^2}{\Delta} \log\left(\frac{1}{\delta}\right)\left(1+o_\delta(1)\right) .$$

On the other hand, the seminal algorithm which is "optimal" for regret minimization is called UCB; it sequentially pulls the arm with the highest "score" (the sum of the empirical average plus some error term) while its decision rule is not satisfied. Although its regret is small, it is not guaranteed that the best arm will be identified in a short time. Here, UCB denotes the classical UCB algorithm, run with confidence parameter $\delta$. It can eliminate one arm when detected as sub-optimal (using the anytime bound used int ETC). $\tau_d$ is actually the number of samples generated before this elimination occurs. Maybe surprisingly, there are no guarantees that this number of samples is actually finite

**Proposition 2** ([2]). *With probability at least $1-\widetilde{\delta}$, the regret of UCB is bounded as*

$$R(\tau_d) \leq \frac{8\sigma^2}{\Delta} \log\left(\frac{1}{\delta}\right)\left(1+o_\delta(1)\right) .$$

*There is no guarantee that $\tau_d$ is uniformly bounded.*

Our family of algorithms interpolates between UCB and ETC by introducing a single parameter $\alpha \in [1,+\infty]$ whose extreme values correspond respectively to focusing solely on regret minimization (i.e., our algorithm identifies with UCB) or best-arm identification (it identifies with ETC). To each value of $\alpha \in [1,+\infty]$ corresponds a different trade-off: Indeed, the bigger the $\alpha$, the bigger the regret but the smaller the decision time.

More precisely, we introduce and study a continuum of algorithms UCB$_\alpha$. For $n \in \mathbb{N}^*$, define $\varepsilon_n = \sqrt{\frac{2\sigma^2}{n}\log(\frac{3\log^2 n}{\delta})}$. For $\alpha, \delta \in \mathbb{R}^+$, UCB$_\alpha$ allocates the user at the population with the highest score $\mathrm{argmax}_i\, \hat{r}^i_{n^i} + \alpha\varepsilon_{n^i}$. It returns an arm $i \in \{A, B\}$ if it dominates the other arm $j$ in the sense that $\hat{r}^i_{n^i} - \varepsilon_{n^i} \geq \hat{r}^j_{n^j} + \varepsilon_{n^j}$.

By construction, the UCB$_\alpha$ algorithm will keep both indexes around the same level. But due to the factor $\alpha > 1$, the intervals of decision with width $\varepsilon_{n^{\mathcal{A}}}$ and $\varepsilon_{n^{\mathcal{B}}}$ (without the factor $\alpha$ that is only used in the sampling policy, not in the decision rule) will eventually become disjoint. Thus, while behaving like a UCB-type algorithm to minimize regret, UCB$_\alpha$ can still return the identity of the best arm. The following theorem makes precise this trade-off between time of decision and regret.

**Algorithm 1** UCB$_\alpha$

```
Input: $\alpha, \delta$
repeat over n
    for each population $i \in \{\mathcal{A}, \mathcal{B}\}$ do
        $\varepsilon^i_n = \sqrt{\frac{2\sigma^2}{n^i}\log(\frac{3\log^2 n^i}{\delta})}$
    end for
    Assign next user to population
        $i_n = \mathrm{argmax}_{i\in\{\mathcal{A},\mathcal{B}\}}\, \hat{r}^i_n + \alpha\varepsilon^i_n$
    $i^* = \mathrm{argmax}_{i\in\{\mathcal{A},\mathcal{B}\}}\, \hat{r}^i_n$
until $\hat{r}^{i^*}_n - \varepsilon^{i^*}_n > \hat{r}^j_n + \varepsilon^j_n$ for $j \neq i$
return $i^*$
```

**Theorem 3.** *With probability greater than $1-\widetilde{\delta}$, UCB$_\alpha$ has a regret $R(\tau_d)$ at its time of decision satisfying*

$$R(\tau_d) \leq \left(\frac{8\sigma^2}{\Delta}c_\alpha + \Delta\right)\log\left(\frac{1}{\delta}\right)\left(1+o_\delta(1)\right) .$$

*where the constant $c_\alpha \in [1,7]$ is defined by*

$$c_\alpha = \min\left\{\frac{(\alpha+1)^2}{4}, \frac{4\alpha^2}{(\alpha-1)^2}\right\} \text{ and } c_1 = 1\ ;\ c_\infty = 4\,.$$

*On that event, the time of decision satisfies $\tau_d \leq T_\delta$ with*

$$T_\delta \leq \frac{\alpha^2+1}{(\alpha-1)^2}\left(\frac{16\sigma^2}{\Delta^2}c_\alpha + 1\right)\log\left(\frac{1}{\delta}\right)\left(1+o_\delta(1)\right) .$$

When $\alpha$ goes to 1, the leading term of the regret goes to $8\sigma^2\log(\frac{1}{\delta})/\Delta$ but the time of decision becomes infinite. When $\alpha \to \infty$, the leading term becomes $32\sigma^2\log(\frac{1}{\delta})/\Delta$ and the time of decision is of order $64\sigma^2\log(\frac{1}{\delta})/\Delta^2$. The fact that the constant $c_\alpha$ is not monotonic in $\alpha$, is an artifact of the proof, i.e., a byproduct of two different analyses of the same problem (for either small or large values of $\alpha$). Indeed, Figure 1 actually indicates that regret of UCB$_\alpha$ is certainly monotonic with respect to $\alpha$ as expected.

ETC has a regret bounded by an expression of order $\frac{16\sigma^2}{\Delta}\log\frac{1}{\delta}$. This is twice bigger than the regret of UCB$_\alpha$ for small $\alpha$. These are only one-sided bounds and do not allow to conclude on which algorithm gets a lower regret, but experiments show that UCB$_\alpha$ for small $\alpha$ indeed presents an advantage in terms of regret versus ETC, at the cost of a higher decision time. See Figure 1.

When $\alpha$ goes to infinity, the exploration term is dominant for UCB$_\alpha$, and it will always pull the least pulled arm. As a consequence, it becomes a variant of ETC with a sub-optimal decision criterion. We denote it by ETC'. It allocates to the populations $\mathcal{A}$ and $\mathcal{B}$ uniformly and selects $\mathcal{A}$ if $\hat{r}^{\mathcal{A}}_n - \hat{r}^{\mathcal{B}}_n \geq \sqrt{\frac{8\sigma^2}{n}\log\left(\frac{3\log^2(n)}{\delta}\right)}$. Its confidence interval width is $\sqrt{2}$ times larger than the one of ETC because of the different concentration arguments used in designing the algorithms: ETC uses a concentration lemma on the difference $\hat{r}_n^{\mathcal{A}} - \hat{r}_n^{\mathcal{B}}$, while UCB$_\alpha$ (and its limit ETC') deals separately with arm $\mathcal{A}$ and $\mathcal{B}$ since their number of samples might be different. Note that the difference between ETC and ETC' is a specificity of the two-armed bandit case, as the generalization of ETC to more than two arms [19] considers per-arm confidence intervals.

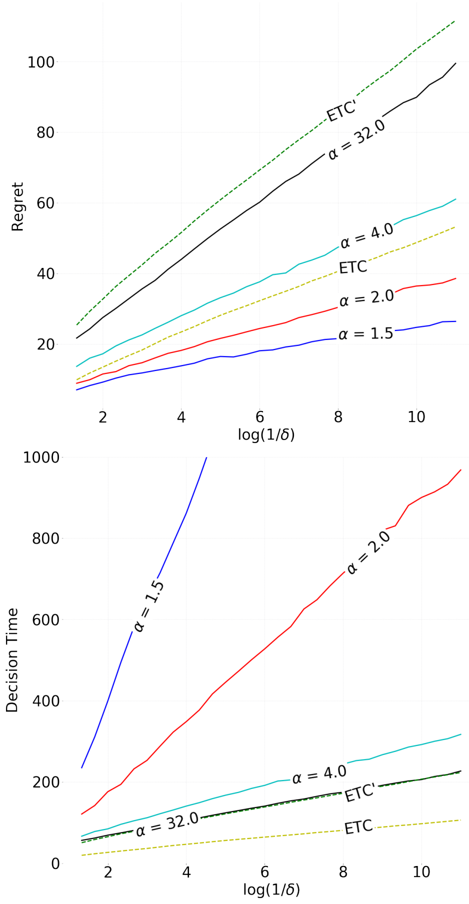

Figure 1: **Comparison of ETC and UCB$_\alpha$. Left: regret at selection. Right: time of selection.** The curves average 1000 experiments with two Gaussian arms of means 0 and 1 and variance 1.

On the other hand, UCB$_\alpha$ generalizes immediately when the number of populations is greater than 2. In that case, UCB$_\alpha$ samples arms as usual but successively eliminates them by looking at the anytime bound on their empirical mean (kind of similarly to LIL-UCB [9]). Then $\tau_d$ corresponds to the time where $K-1$ arms have been eliminated and the algorithm can output a candidate for the best arm. We assume that the index of the optimal population is 1 and we denote by $\Delta_k = r^1 - r^k$ the gap associated to the subobtimal population $k$. Then we can derive the following corollary from Theorem 3.

**Corollary 4.** *With $K$ different arms, and with probability greater than $1-\widetilde{\delta}$, UCB$_\alpha$ has a decision time smaller than*

$$\left(\frac{(\alpha+1)^2}{(\alpha-1)^2}\left(\frac{8\sigma^2}{\Delta_{\min}^2}c_\alpha+1\right)+\sum_{k=2}^{K}\frac{8\sigma^2}{\Delta_k^2}c_\alpha+K\right) \times \log\left(\frac{K}{\delta}\right)\left(1+o_\delta(1)\right)$$

*and its regret $R(\tau_d)$ satisfies*

$$R(\tau_d) \leq \sum_{k=2}^{K}\left(\frac{8\sigma^2}{\Delta_k}c_\alpha+\Delta\right)\log\left(\frac{K}{\delta}\right)\left(1+o_\delta(1)\right).$$

With $K>2$ population, ETC would stop sampling population $k$ as soon as there exists another population $i$ such that $\hat{r}_n^i - \hat{r}_n^k \geq 4\sqrt{\frac{\sigma^2}{n}\log\left(\frac{3K\log^2(n)}{\delta}\right)}$. As a consequence, its decision time will be upper-bounded by

$$\left(\frac{32\sigma^2}{\Delta_{\min}^2}+\sum_{k=2}^{K}\frac{32\sigma^2}{\Delta_k^2}\right)\log(\frac{K}{\delta})(1+o_\delta(1))$$

and its regret smaller than

$$\left(\sum_{k=2}^{K}\frac{32\sigma^2}{\Delta^2}\log(\frac{K}{\delta})\right)(1+o_\delta(1)).$$

As a consequence, even if UCB$_\alpha$ interpolates between UCB and ETC in term of regret, it actually outperforms ETC in terms of decision time when $\alpha\to\infty$.

### 1.1 How does inflating the exploration term lead to a finite decision time? A proof sketch

We consider an algorithm which pulls $\arg\max_{i\in\{\mathcal{A},\mathcal{B}\}} \hat{r}_{n^i}^i + \alpha\varepsilon_{n^i}$ with $\alpha>1$, and stops if $|r_{n^{\mathcal{A}}}^{\mathcal{A}} - r_{n^{\mathcal{B}}}^{\mathcal{B}}| > \varepsilon_{n^{\mathcal{A}}} + \varepsilon_{n^{\mathcal{B}}}$ and returns the arm with highest mean at that point.

Recall that $\varepsilon_n$ is a quantity close to $1/\sqrt{n}$, up to logarithmic terms in $n$ and multiplicative constants, hence $1/\varepsilon_n^2 \approx cn$ for some constant $c$. If we can prove that as long as no decision is taken, $1/\varepsilon_{n^{\mathcal{A}}}^2$ and $1/\varepsilon_{n^{\mathcal{B}}}^2$ are bounded from above, then we obtain an upper bound on the time $t = n_{\mathcal{A}} + n_{\mathcal{B}}$ before a decision is taken.

Suppose that the best arm is $\mathcal{A}$. Bounding $n_{\mathcal{B}}$ is done through classical bandit arguments: since $\mathcal{B}$ is the worse arm, a UCB-type algorithm does not pull it often. The challenge is to show that our algorithm also controls $n_{\mathcal{A}}$.

We can make use of a concentration result of the form: with probability $1-\tilde{\delta}$, $\hat{r}^{\mathcal{A}}(n^{\mathcal{A}}) + \varepsilon_{n^{\mathcal{A}}} \geq r^{\mathcal{A}}$ and $\hat{r}^{\mathcal{B}}(n^{\mathcal{B}}) - \varepsilon_{n^{\mathcal{B}}} \leq r^{\mathcal{B}}$. The decision criterion ensures that if a decision is taken and these concentration inequalities hold, then the arm returned is the correct one. Indeed under this concentration event, $\hat{r}^{\mathcal{B}}(n^{\mathcal{B}}) - \hat{r}^{\mathcal{A}}(n^{\mathcal{A}}) \leq r^{\mathcal{B}} - r^{\mathcal{A}} + \varepsilon_{n^{\mathcal{B}}} + \varepsilon_{n^{\mathcal{A}}}$ , which is strictly smaller than $\varepsilon_{n^{\mathcal{B}}} + \varepsilon_{n^{\mathcal{A}}}$ . Hence $\mathcal{B}$ cannot be returned: if the algorithm stops, it is correct.

The algorithm will keep both indexes roughly equal, hence $\hat{r}^{\mathcal{A}}_{n^{\mathcal{A}}} + \alpha\varepsilon_{n^{\mathcal{A}}} \approx \hat{r}^{\mathcal{B}}_{n^{\mathcal{B}}} + \alpha\varepsilon_{n^{\mathcal{B}}}$ and the "pulling" confidence intervals with width $\alpha\varepsilon_n$ will never get disjoint. But as $\varepsilon_{n^{\mathcal{A}}}$ and $\varepsilon_{n^{\mathcal{B}}}$ get small, the smaller "decision" confidence intervals with width $\varepsilon_n$ will eventually separate, as seen in Figure 2.

More formally, if $\mathcal{A}$ is pulled and no decision was taken yet, then the index of $\mathcal{A}$ is big, $\hat{r}^{\mathcal{A}}(n^{\mathcal{A}}) + \alpha\varepsilon_{n^{\mathcal{A}}} > \hat{r}^{\mathcal{B}}(n^{\mathcal{B}}) + \alpha\varepsilon_{n^{\mathcal{B}}}$, but not so big that the algorithm stops, i.e. $\hat{r}^{\mathcal{A}}(n^{\mathcal{A}}) - \varepsilon_{n^{\mathcal{A}}} \leq \hat{r}^{\mathcal{B}}(n^{\mathcal{B}}) + \varepsilon_{n^{\mathcal{B}}}$. By combining the two, we obtain the relation $n_{\mathcal{A}} \propto \frac{1}{\varepsilon^2_{n^{\mathcal{A}}}} \leq \frac{(\alpha+1)^2}{(\alpha-1)^2}\frac{1}{\varepsilon^2_{n^{B}}} \propto \frac{(\alpha+1)^2}{(\alpha-1)^2} n^{\mathcal{B}}$ , where $\propto$ is to be read as the informal statement that the quantities are roughly proportional.

To sum-up the idea of the proof: $\mathcal{B}$ will not be pulled much since it is the worst arm and we employ an UCB-type algorithm. $\mathcal{A}$ will be pulled less than a factor depending on $\alpha$ times $\mathcal{B}$. Thus as long as no decision is taken, $t = n^{\mathcal{A}} + n^{\mathcal{B}}$ is bounded, by a quantity $T_\delta$. We conclude that the decision time is smaller than $T_\delta$.

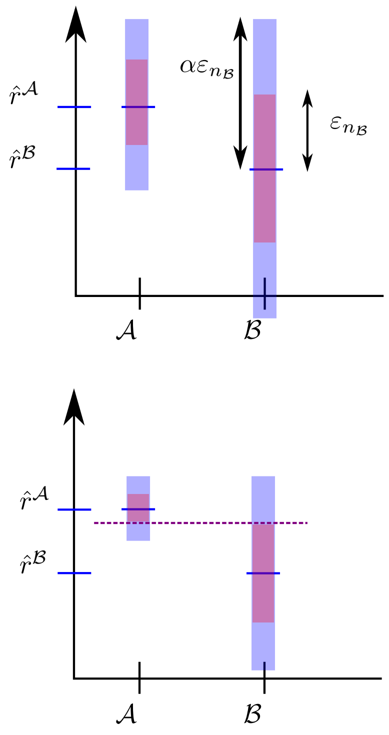

Figure 2: **Confidence intervals used in UCB$_\alpha$**. Top: Before the time of decision, the indexes $\hat{r}(n) + \alpha\varepsilon_n$ are aligned, the tighter $\varepsilon_n$ intervals also overlap. Bottom: time of selection: the tight $\varepsilon_n$ confidence intervals become disjoint.

## 2 Extensions to non-iid settings

When an internet platform wants to A/B test two versions of its website, purchases from customers that are buying recurrently can not be considered as independent. When a technical A/B test is run, it is also usual to split servers in two populations to A/B test the business impact of the latency of a new code version. In this case, observations from the same server are not independent. This setting is not restricted to online marketing. Clinical trials classically measure the survival time or quality of life of patients over time and adaptive testing is a key challenge in this context, however it is so far either heuristic [21] or under the restrictive assumption to observe the rewards before the next decision [3, 8]. More recently, [17] studied a similar setting of reward arriving over time, however with a different objective, namely finding an adaptive way to stop the test for a patient taking into account a cost of testing.

The setting of multi-armed bandit presented in the previous section has been applied to A/B tests [13, 23, 24] with independent rewards. The goal of this section is to show that we can extend algorithms presented in the first section in a more complex setting and provide a framework to practitioners for interpolating between best arm identification and regret minimization in other settings than the iid case.

To model theses aspects, we show that decisions of the multi-armed bandit algorithm can be taken at a unit level (e.g. users, patients, servers...). Once allocated to a population $\mathcal{A}$ or $\mathcal{B}$, a unit $u$ interacts with the system during the whole A/B test . When time increases, the system gathers more and more signal on a unit arrived early in the A/B test. We will also assume in order to be able to take a causal decision at the end of the A/B test that units already exposed to one treatment can not be switched from population. A unit stays in the same population during the whole A/B test. Intuitively, the system will estimate the performance of one technology by averaging its performance on the different units.

### 2.1 Notations

We need to differentiate the units (e.g. users) randomly assigned to populations $\mathcal{A}$ and $\mathcal{B}$ from their associated rewards. In the iid setting, the reward $r^{\mathcal{A}}_u$ associated to a unit $u$ in population $\mathcal{A}$ was assumed to be observed instantly. Now, we assume that, for each unit $u$ we have been seeing so far, we observe noisy version of this reward over time and $r^{\mathcal{A}}_u$ is only the unknown expectation of this process. Population-specific notation is symmetric, thus, for the sake of readability, we only detail notations for the control $\mathcal{A}$ and assume the corresponding one for the treatment $\mathcal{B}$.

More formally, we assume the units $u$ are i.i.d. samples from an unknown distribution and arriving in the test dynamically over time. To each unit $u$ is associated an unknown reward $r^{\mathcal{A}}_u$ which is an i.i.d. sub-gaussian r.v. with expected value

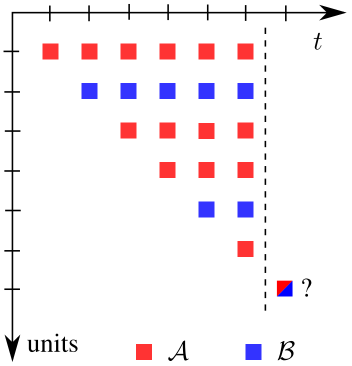

Figure 3: **The unit allocation problem**. At each stage $t \in \mathbb{N}$, a new unit is allocated to either population $\mathcal{A}$ or $\mathcal{B}$ and a sample from every unit arrived before $t$ is observed.

$r^{\mathcal{A}}$ and variance $\sigma_r^2$, as well as an arrival time $\mathcal{T}_u$. $\mathcal{A}(t)$ denotes the set of units in $\mathcal{A}$ of cardinality $n_t^{\mathcal{A}}$ at time $t$. Then, given a unit $u$ and starting at time $\mathcal{T}_u$, we observe over time random outcomes $r_{u,t}^{\mathcal{A}} = r_u^{\mathcal{A}} + \varepsilon_{u,t}$ where $\varepsilon_{u,t}$ is a zero-mean sub-gaussian variable with variance $\sigma_\varepsilon^2$. At time $t$, the unit $u$ has generated $t - \mathcal{T}_u + 1$ outcomes $r_{u,t_u}^{\mathcal{A}} \cdots r_{u,t}^{\mathcal{A}}$.

At each time step, the algorithm has access to all the rewards generated by all the users already present in the A/B test and the precision on $r_u^{\mathcal{A}}$ will increase with time as more samples $r_{u,t}^{\mathcal{A}}$ are gathered.

In this setting, a natural estimator to consider is

$$\hat{r}_t^{\mathcal{A}} := \frac{1}{n_t^{\mathcal{A}}} \sum_{u \in \mathcal{A}(t)} \frac{1}{t - \mathcal{T}_u + 1} \sum_{s=\mathcal{T}_u}^{t} r_{u,s}^{\mathcal{A}}$$

We call this estimator the *Mean of means* estimator. With this estimator, we can design algorithms that can reach a trade-off between regret minimization and best arm identification and see how this regret depend on $\sigma_r^2$ and $\sigma_\varepsilon^2$. We could have considered other estimators such that the *Total Mean* estimator defined as

$$\hat{R}_t^{\mathcal{A}} := \frac{1}{\sum_{u \in \mathcal{A}(t)} t - \mathcal{T}_u + 1} \sum_{u \in \mathcal{A}(t)} \sum_{s=\mathcal{T}_u}^{t} r_{u,s}^{\mathcal{A}}$$

which is also an unbiased estimator of $r$. Yet unfortunately, this estimator puts more weights on older units than on the more recent one. In the case where all units have more or less the same noise, this is not really an issue (but this is more or less the only case where the total mean estimator has good behavior). Unfortunately, with adaptive sampling algorithm (such as UCB), then it could be the case that a new unit is allocated to a population after an exponential long time. In that case, the different weights put on different units can be of different order of magnitude, preventing fast convergence of the estimate to the estimated mean.

A second motivation for choosing the *Mean of means* estimator is because it opens doors for generalizing to more complex models on the stochastic processes underlying the units behavior. Indeed, it can be seen as an average over the units of the per-unit stochastic process expected values. Here, we assume the random outcomes $r_{u,t}^i$ are i.i.d. with expectation $r_u^i$ (technically our results hold for martingale difference sequence) . Then the technical part shows how to combine concentration results on each of the units to derive bandit algorithms with good properties. With a different model on the per-user random outcomes (as. mean reverting process or cyclic process), our proof techniques could be used as long as it is possible to construct an estimator of the expectation that concentrate well enough.

Precise statements and proofs of results presented in this section can be found in Appendix C.

### 2.2 ETC

The ETC algorithm allocates units alternatively to $\mathcal{A}$ and $\mathcal{B}$, choosing possibly the first population at random. To simplify the analysis, we are actually going to assume that 2 units arrive at each stage, one of each being allocated to each population. So that if ETC stops at stage $t \in \mathbb{N}$, then both populations have $n = t$ units, and regret is $n\Delta$.

The stopping rule criterion of ETC is simply the following:

$$|\hat{r}_n^{\mathcal{A}} - \hat{r}_n^{\mathcal{B}}| > \sqrt{\frac{4\left(\sigma_r^2 + \frac{\sigma_\varepsilon^2 \log(en)}{n}\right)}{n} \log\left(\frac{\pi^2 n^2}{3\delta}\right)}$$

**Theorem 5.** *With probability at least $1 - \delta$, the ETC algorithm outputs the best population and stops after having allocating a total of at most $\tau_d \leq T_\delta$ units, where.*

$$T_\delta = \frac{32\sigma_r^2}{\Delta^2} \log(\frac{1}{\delta})(1 + o_\delta(1)) + \frac{\sigma_\varepsilon^2}{\sigma_r^2} \log\log(\frac{1}{\delta})(1 + o_\delta(1)) \,.$$

*Its regret at $\tau_\delta$ is equal to $\frac{\Delta}{2}\tau_\delta$.*

### 2.3 UCB-MM

The variant of UCB we consider is defined with respect to the following index of performance of population $i \in \{\mathcal{A}, \mathcal{B}\}$ defined by

$$\hat{r}_t^i + \sqrt{\frac{2\left(\sigma_r^2 + \frac{\sigma_\varepsilon^2 \log(en_t^i)}{n_t^i}\right)}{n_t^i} \log\left(\frac{(4n_t^i)^4}{2\delta} \max\{1, \frac{n_t^i \sigma_r^2}{\sigma_\varepsilon^2}\}\right)} \tag{1}$$

Using this index, UCB-MM is described in Algorithm 2. We mention here that having random numbers of units with random numbers of occurrences prevents us for deriving the more or less standard concentration inequalities derived for the other algorithm. Indeed, it actually requires to combine

**Algorithm 2** UCB-MM$_\alpha$

```
Input: $\alpha, \delta$
repeat over t
    for each population $i \in \{\mathcal{A}, \mathcal{B}\}$ do
        $\varepsilon_t^i = \sqrt{\frac{2\left(\sigma_r^2 + \frac{\sigma_\varepsilon^2 \log(en_t^i)}{n_t^i}\right)}{n_t^i}}$
                $\times \sqrt{\log\left(\frac{(4n_t^i)^4}{\delta} \max\{1, \frac{n_t^i \sigma_r^2}{\sigma_\varepsilon^2}\}\right)}$
    end for
    Assign next user to population
        $i_t = \mathrm{argmax}_{i \in \{\mathcal{A}, \mathcal{B}\}} \hat{r}_t^i + \alpha \varepsilon_t^i$
    $i^* = \mathrm{argmax}_{i \in \{\mathcal{A}, \mathcal{B}\}} \hat{r}_t^i$
until $\hat{r}_t^{i^*} - \varepsilon_t^{i^*} > \hat{r}_t^j + \varepsilon_t^j$ for $j \neq i$
return $i^*$
```

several types of different inequalities, which explains the non-standard term in the $\sqrt{\log(\cdot)}$ part of the index.

**Theorem 6.** *With probability at least $1 - \widetilde{\delta}$, the regret of UCB-MM is bounded at stage $t_d$ as*

$$R(\tau_d) \leq \frac{8\sigma_r^2}{\Delta} \log(\frac{1}{\delta})(1 + o_\delta(1)) + \frac{\sigma_r^2}{\sigma_\varepsilon^2} \Delta \log\log(\frac{1}{\delta})(1 + o_\delta(1)) \,.$$

*There is no guarantee that $\tau_d$ is uniformly bounded.*

We emphasize here that the dependency of the total regret with respect to the noise variance $\sigma_\varepsilon^2$ is negligible compared to its dependency with respect to the variance of the unit performance $\sigma_r^2$. Indeed, regret has a $\log \frac{1}{\delta}$ factor in front of $\sigma_r^2$ (a term which is unavoidable, even without extra noise, i.e., if $\sigma_\varepsilon^2 = 0$). On the other hand, the multiplicative factor of $\sigma_\varepsilon^2$ is only double logarithmic, in $\log\log\frac{1}{\delta}$. Moreover, the additional number of units required to find the best population is, asymptotically, independent of $\Delta$, the proximity measure of the two populations.

In the index definition of UCB-MM (Equation (1)), letting $\sigma_\varepsilon^2$ goes to 0 gives a void bound (the index is basically always $+\infty$). This is an artefact of the proof needed for having only an extra $\log\log(\cdot)$ term. It is also possible to use the following alternative error term for UCB-MM

$$\sqrt{\frac{2\sigma_r^2}{n_t^i} \log\left(\frac{9\log^2(n_t^i)}{\delta}\right)} + \sqrt{\frac{2\sigma_\varepsilon^2 \log(en_t^i)}{(n_t^i)^2} \log\left(\frac{(4n_t^i)^4}{\delta}\right)}$$

This error term converges, as $\sigma_\varepsilon^2$ goes to 0, to the usual error term of UCB. Unfortunately, the regret dependency in $\sigma_\varepsilon^2$ deteriorates as it scales with

$$\left(\frac{8\sigma_r^2}{\Delta} \log\frac{1}{\delta} + \frac{\sigma_r^2}{\sigma_\varepsilon^2} \Delta \sqrt{\log\frac{1}{\delta}}\right)(1 + o_\delta(1))$$

thus with a $\sqrt{\log(\cdot)}$ extra term instead of a $\log\log(\cdot)$ one.

### 2.3.1 Interpolating between regret minimization and best arm identification

As in the iid case, it is possible, with unit, to define UCB-MM$_\alpha$ to interpolate between UCB-MM and ETC-MM by multiplying the error term of UCB-MM by a factor $\alpha \in [1, +\infty]$.

**Theorem 7.** *With probability at least $1 - \widetilde{\delta}$, the regret of UCB-MM$_\alpha$ is bounded at stage $T$ as*

$$R(\tau_d) \leq \left(\frac{8\sigma_r^2}{\Delta} c_\alpha + \Delta\right) \log(\frac{1}{\delta})(1 + o_\delta(1)) + \frac{\sigma_r^2}{\sigma_\varepsilon^2} \Delta \log\log(\frac{1}{\delta})(1 + o_\delta(1)) \,.$$

*Moreover, the time of decision is upper-bounded by*

$$\tau_d \leq \frac{\alpha^2 + 1}{(\alpha - 1)^2} \left(\frac{16\sigma_r^2}{\Delta} c_\alpha + 1\right) \log\left(\frac{1}{\delta}\right)\left(1 + o_\delta(1)\right) + 2\frac{\sigma_r^2}{\sigma_\varepsilon^2} \Delta \log\log(\frac{1}{\delta})(1 + o_\delta(1)) \,.$$

The proof of this result proceeds as in the iid case (see the sketch of proof in section 1.1), assuming that we can provide a suitable concentration inequality to bound the deviation of the mean-of-means estimator. The main difference is the concentration arguments used. The analysis is adaptive in that it gives a bound on the decision time for any model, as long as we are able to provide concentration inequalities for the population means. The concentration inequality is used to obtain a bound on the number of allocations to the sub-optimal population, then the inflated confidence intervals mechanically provide a bound on the number of pulls of the best population with respect to the sub-optimal one.

As in the iid case, we have stated our results for $K = 2$ populations, but it's straightforward to generalize them for $K > 2$ different populations.

**Practical remark.** In practice, attributing users dynamically to populations could be hard to handle in production (for instance, the population of the user must be stored to interact with him when he is coming back on the platform..). This is why we also provide another anytime bound in appendix in a simpler setting where we assume that all the units are already present at the beginning of the test, thus allowing to allocate them to populations using a simple hash on their identifiers. Based on this bound, the practitioner can stop the test as early as possible such that the decision is statically sufficient. However, this bound can not help to do a tradeoff between regret minimization and best arm identification.

On the other hand, we assume that the reward and noise were subGaussian random variable, with known variance (proxy) $\sigma_r^2$ and $\sigma_\varepsilon^2$. Our results and techniques can be generalized to the case where the random variables $r_t^i$ and $\varepsilon_{n,t}$ are bounded (say, in $[0, 1]$) with unknown variance. One just need to use empirical Bernstein concentration inequalities [16].

## 3 Experiments

On a simple iid setting, we show the performance proved in Section 1 on Figure 1. There are two Gaussian arms with same variance 1 and means 0 and 1. The two graphs on the figure show the decision time and the regret at the time of decision of several algorithms for a range of values of $\log(1/\delta)$. The algorithms shown are ETC, four instances of UCB$_\alpha$ for alpha in $(1.5, 2, 4, 32)$ and ETC', the variant of ETC to which UCB$_\alpha$ tends to when $\alpha \to \infty$. We highlight a few conclusions from these plots, which are all in agreement with the theoretical results.

- The algorithm with lowest decision time is ETC, the algorithm with lowest regret is UCB$_\alpha$ with small $\alpha$.
- For $\alpha \geq 4$, UCB$_\alpha$ has lower regret and higher decision time than ETC', but is worse than ETC on both criteria.
- For $\alpha$ equal to 1.5 or 2, UCB$_\alpha$ has lower regret and higher decision time than ETC.

UCB$_\alpha$ is seen empirically to realize a trade-off between its two limiting algorithms UCB and ETC', and there is an interval for $\alpha$ in which UCB$_\alpha$ is a trade-off between UCB and ETC. The numerical relations between the bounds can also be verified: the regret of UCB$_{1.5}$ is almost twice smaller than the regret of ETC, which is twice smaller than the regret of ETC'.

We then show on Figure 4 empirical results on the unit setting presented in Section 2. At each time step, a user arrives and then generate a reward according to $r_u$ for every time step until the end of the game. In our simulation, $r_u$ is sampled from a normal distribution with mean 0 for population $\mathcal{A}$ (respectively with mean 1 for population $\mathcal{B}$) and variance $\sigma_r^2$ equal to 1. The noise $\epsilon$ at each time step is also Gaussian of mean 0 and variance $\sigma_\epsilon^2$ equal to 1. The data can be seen as a triangular matrix (cf Fig. 3). We compare performances between UCB$_\alpha$ for different values of $\alpha$ and ETC' in terms of regret and times of decision. We see that we are able to reproduce what we observed in the iid setting in the unit setting. We observe again a factor 4 between the regret of ETC' and UCB. With UCB$_\alpha$, we can realize a trade-off between ETC' and UCB, both in terms of regret and decision time.

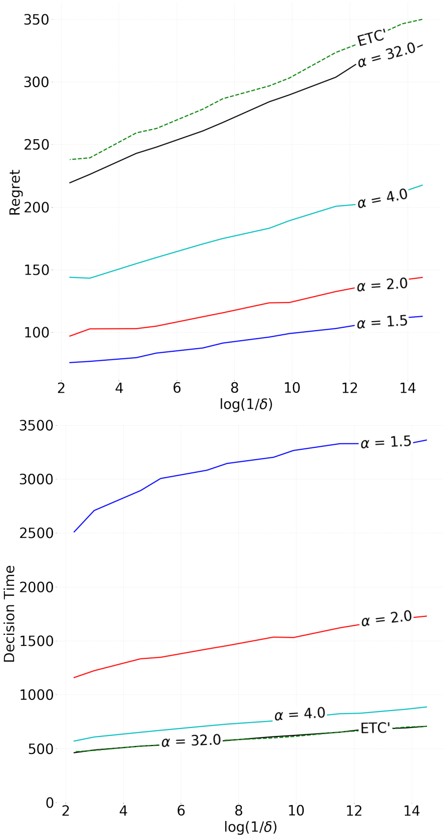

Figure 4: **First: Regret in the unit setting. Second: Decision time for the unit setting.** The curves are averages of 1000 experiments with two Gaussian arms with means 0 and 1 and $\sigma_r = 1$, $\sigma_\epsilon = 1$.

## 4 Conclusion

We studied A/B tests in the fixed confidence setting from the two perspectives of regret minimization and best arms identification. We introduced a class of algorithms that optimizes at the same time both objectives. It interpolates between optimal algorithms designed for each case. Our study also shed light on an effect often seen in practice: the UCB algorithm not only minimizes regret but also identifies the best arm in finite time, as soon as its exploration term is slightly inflated. This inflation is nearly always present in practice. Indeed, for UCB to be a valid algorithm for a noise and the confidence interval not to be bigger than necessary, it would need to be run with a constant matching exactly the unknown sub-Gaussian norm of the noise.

We extended our study to a non-iid setting by deriving adapted concentration results for the mean-of-means estimator, again obtaining algorithms which interpolate between the two objectives. We would however like to warn practitioners on an intensive use of bandit algorithms for A/B tests. Even if data are collected through time, it can prove to be difficult to define a time of arrival for units not correlated with the data to ensure the iid assumption holds as bandit algorithms that neglect this aspect could behave quite poorly. Handling such problem would require to model the unit stochastic process *conditionally* to the time of arrival which would be a use case for generalizing our results to more complex stochastic processes for unit modelling.

## 5 Aknowledgement

The fourth author has benefited from the support of the FMJH Program Gaspard Monge in optimization and operations research (supported in part by EDF), from the Labex LMH and from the CNRS through the PEPS program.

## References

[1] J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, 2010.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Journal of Machine Learning Research*, 47(2-3), 2002.

[3] D. A. Berry and S. G. Eick. Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine*, 14(3), 1995.

[4] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412, 2011.

[5] A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, 2016.

[6] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(Jun), 2006.

[7] A. Garivier, T. Lattimore, and E. Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, 2016.

[8] F. Hu and W. F. Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*. John Wiley & Sons, Inc., Apr 2006.

[9] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, 2014.

[10] R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17. ACM, 2017.

[11] Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, 2013.

[12] M. N. Katehakis and H. Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19), 1995.

[13] E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of A/B Testing. In *Conference on Learning Theory*, Proceedings of The 27th Conference on Learning Theory, 2014.

[14] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1), 2016.

[15] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun), 2004.

[16] A. Maurer and M. Pontil. Empirical Bernstein Bounds and Sample-Variance Penalization. *Conference on Learning Theory*, 2009.

[17] D. M. Negoescu, K. Bimpikis, M. L. Brandeau, and D. A. Iancu. Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science*, 64(8), 2018.

[18] L. Pekelis, D. Walsh, and R. Johari. The new stats engine. *Internet. Retrieved December*, 6, 2015.

[19] V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 2013.

[20] V. Perchet, P. Rigollet, S. Chassang, E. Snowberg, et al. Batched bandit problems. *The Annals of Statistics*, 44 (2), 2016.

[21] W. H. Press. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proc Natl Acad Sci U S A*, 106 (52), Dec 2009.

[22] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4), 1933.

[23] F. Yang, A. Ramdas, K. G. Jamieson, and M. J. Wainwright. A framework for multi-a(rmed)/b(andit) testing with online fdr control. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[24] S. Zhao, E. Zhou, A. Sabharwal, and S. Ermon. Adaptive concentration inequalities for sequential decision problems. In *Advances in Neural Information Processing Systems*. 2016.