Thresholding Bandits with Augmented UCB

Author names withheld

Abstract

To be written

1 Introduction

In this paper we study a specific combinatorial pure exploration problem called thresholding bandit problem in the stochastic multi-armed bandit setting. In the stochastic multi-armed bandit setting a learning agent is required to choose from a set of decisions or arms at every round. The agent is then presented with a reward for that round, which is an independent draw from a stationary distribution specific to the arm selected. The agent, however, does not know the mean of the distributions associated with each arm, denoted by r_i, including the optimal arm which will give it the best reward, denoted by r^*. The agent attempts to make arm choices that will maximize some performance measure by keeping track of the reward that has been gathered from previous selections of the arm, for each arm. This is called the estimated mean reward of an arm denoted by $\hat{r}_i$. The bandit problem can be conceptualized as a sequential decision making process where the agent is at each round presented with an *exploration-exploitation dilemma*. The agent could pull the arm which has the highest observed mean reward till now (exploitation) or to explore other arms, with the prospect of finding superior performance which was previously unobserved (exploration).

In the pure exploration thresholding bandit setup the goal is different than minimizing the cumulative regret, that is the total loss suffered by the learner for not selecting the optimal arm throughout the time horizon T. Here the learning algorithm is provided with a threshold τ and it has to output all such arms i whose mean of reward distribution r_i is above τ after T rounds. This is a specific instance of combinatorial pure exploration where the learning algorithm can explore as much as possible given a fixed horizon T and not be concerned with the usual exploration-exploitation dilemma. Let A be the set of all arms. Formally we can define a set $S_\tau = \{i \in A : r_i \geq \tau\}$ and the complementary set $S_\tau^C = \{i \in A : r_i < \tau\}$. Also we define $\hat{S}_\tau = \hat{S}_\tau(T) \subset A$ and its complementary set $\hat{S}_\tau^C$ as the recommendation of the learning algorithm after T rounds. Given such sets exists, the performance of the learning agent is measured by how much accuracy it can discriminate between S_τ and S_τ^C after time horizon T. The loss $\mathcal{L}$ is defined as:-

$$\mathcal{L}(T) = I\big(\{S_\tau \cap \hat{S}_\tau^C \neq \emptyset\} \cup \{\hat{S}_\tau \cap S_\tau^C \neq \emptyset\}\big)$$

The goal of the learning agent is to minimize $\mathcal{L}(T)$. So, the expected loss after T rounds is,

$$\mathbb{E}[\mathcal{L}(T)] = \mathbb{P}\big(\{S_\tau \cap \hat{S}_\tau^C \neq \emptyset\} \cup \{\hat{S}_\tau \cap S_\tau^C \neq \emptyset\}\big)$$

which we can say is the probability of making mistake, that is whether the learning agent at the end of round T rejects arms from S_τ or accepts arms from S_τ^C in its final recommendation. Also, we are looking at an anytime algorithm, so the knowledge of T may not be known to the learner.

2 Motivation

The thresholding bandit problem (TBP) has several relevant industrial applications. In some cases the TBP problem is more relevant than the variants of TopM problem (identifying the best M arms from K given arms).

1. *Product Selection:* A company wants to introduce a new product in market and there is a clear separation of the test phase from the commercialization phase. In this case the company tries to minimize the loss it might incur in the commercialization phase by testing as much as possible in the test phase. So from the several variants of the product that are in the test phase the learning agent must suggest the product variant(s) that are above a particular threshold τ at the end of the test phase that have the highest probability of minimizing loss in the commercialization phase. A similar problem has been discussed for single best product variant identification without threshold in [Bubeck *et al.*, 2011].

2. *Mobile Phone Channel Allocation:* Another similar problem as above concerns channel allocation for mobile phone communications ([Audibert *et al.*, 2009]). Here there is a clear separation between the allocation phase and communication phase whereby in the allocation phase a learning algorithm has to explore as many channels as possible to suggest the best possible set of channel(s) that are above a particular threshold τ. The threshold depends on the subscription level of the customer. With higher subscription the customer is allowed better channel(s) with the τ set high. Each evaluation of a channel is noisy and the learning algorithm must come up with the best possible suggestion within a very small number of attempts.

3. *Anomaly Detection and Classification:* Thresholding bandit can also be used for anomaly detection and classification where we define a cutoff level τ and for any samples above this cutoff gets classified as an anomaly. For further reading we point the reader to section 3 of [Locatelli *et al.*, 2016].

3 Contribution

To be written

4 Related Works and Previous Results

A significant amount of work has been done on the stochastic multi-armed bandit setting regarding minimizing cumulative regret with a single optimal arm. For a survey of such works we refer the reader to [Bubeck and Cesa-Bianchi, 2012]. An early work involving a bandit setup is [Thompson, 1933], where the author deals with the problem of choosing between two treatments to administer on patients who come in sequentially. Following the seminal work of [Robbins, 1952], bandit algorithms have been extensively studied in a variety of applications. From a theoretical standpoint, an asymptotic lower bound for the regret was established in [Lai and Robbins, 1985]. Several other works such as [Auer *et al.*, 2002a], [Audibert and Bubeck, 2009] and [Auer and Ortner, 2010] have shown results for minimizing cumulative regret in stochastic bandit setup whereas works such as [Auer *et al.*, 2002b] have concentrated on adversarial bandit setup.

In the pure exploration setup, a significant amount of research has been done on finding the best arm(s) from a set of arms. The pure exploration setup has been explored in mainly two settings:-

1. Fixed Budget setting: In this setting the learning algorithm has to suggest the best arm(s) within a fixed number of attempts that is given as an input. The objective here is to maximize the probability of returning the best arm(s). One of the foremost papers to deal with single best arm identification is [Audibert *et al.*, 2009] where the authors come up with the algorithm UCBE and Successive Reject(SR) with simple regret guarantees. The relationship between cumulative regret and simple regret is proved in [Bubeck *et al.*, 2011] where the authors prove that minimizing the simple regret necessarily results in maximizing the cumulative regret. In the combinatorial fixed budget setup [Gabillon *et al.*, 2011] come up with Gap-E and Gap-EV algorithm which suggests the best m (given as input) arms at the end of the budget with high probability. Similarly, [Bubeck *et al.*, 2013] comes up with the algorithm Successive Accept Reject(SAR) which is an extension of the SR algorithm. SAR is a round based algorithm whereby at the end of round an arm is either accepted or rejected based on certain conditions till the required top m arms are suggested at the end of the budget with high probability.
2. Fixed Confidence setting: In this setting the the learning algorithm has to suggest the best arm(s) with a fixed (given as input) confidence with as less number of attempts as possible. The single best arm identification has been handled in [Even-Dar *et al.*, 2006] where they come up with an algorithm called Successive Elimination (SE) which comes up with an arm that is ϵ close to the optimal arm. In the combinatorial setup recently [Kalyanakrishnan *et al.*, 2012] have suggested the LUCB algorithm which on termination returns m arms which are atleast ϵ close to the true top m arms with $1-\delta$ probability.

Apart from these two settings some unified approach has also been suggested in [Gabillon *et al.*, 2012] which proposes the algorithms UGapEb and UGapEc which can work in both the above two settings. A similar combinatorial setup was also explored in [Chen *et al.*, 2014] where the authors come up with more similarities and dissimilarities between these two settings in a more general setup. In their work, the learning algorithm, called Combinatorial Successive Accept Reject (CSAR) is similar to SAR with a more general setup. The thresholding bandit problem is a specific instance of the pure exploration setup of [Chen *et al.*, 2014]. In the latest work in [Locatelli *et al.*, 2016] the algorithm Anytime Parameter-Free Thresholding (APT) algorithm comes up with a better anytime guarantee than CSAR for the thresholding bandit problem.

5 Notation Used and Assumptions

In this paper A is the set of all arms and $|A| = K$ denotes the number of arms in the set. Any arm is denoted by i. The average estimated payoff for any arm is denoted by $\hat{r}_i$ whereas the true mean of the distribution from which the rewards are sampled is denoted by r_i. The optimal arm is denoted by $*$. The '*' superscript is used to denote anything related to optimal arm. $\Delta_i = |\tau - r_i|$ and $\hat{\Delta}_i = |\tau - \hat{r}_i|$. Also we define $\Delta_i = r^* - r_i$ and $\hat{\Delta}_i = \hat{r}^* - \hat{r}_i$. In all cases $\min_{i \in A} \Delta_i$ is denoted by Δ. n_i denotes the number of times the arm i has been pulled. ψ denotes the exploration regulatory factor and ρ, ρ_v as arm elimination parameters. $\hat{V}_i = \frac{1}{t} \sum_{t=1}^{n_i} (x_{i,t} - r_i)^2$ denotes the empirical variance and $x_{i,t}$ is the reward obtained at timestep t for arm i. Also σ_i^2 denotes the true variance of the arm.

It is assumed that the distribution from which rewards are sampled are identical and independent 1-sub-Gaussian distributions. Throughout the paper, we assume that the distributions v_i are 1-sub-Gaussian including Gaussian distributions with variance less than 1 and distributions supported on an interval of length less than 2. We will also assume that all rewards are bounded in $[0, 1]$.

6 Augmented UCB

In algorithm 1, hence referred to as AugUCB, we have three exploration parameters, ρ, ρ_v which are the arm elimination parameters and ψ which is the exploration regulatory factor. The threshold τ is also given as an input. AugUCB combines the power of UCB-Improved ([Auer and Ortner, 2010]), APT ([Locatelli *et al.*, 2016]) and SAR ([Gabillon *et al.*, 2011]) or CSAR([Chen *et al.*, 2014]). The main approach is based on UCB-Improved with modifications suited for the thresholding bandit problem. The active set B_0 is initialized with

Algorithm 1 AugmentedUCB

Input: Time horizon T, exploration parameters ρ, ρ_v and ψ, threshold τ.

Initialization: Set $B_0 := A$, $M = \left\lfloor \frac{1}{2}\log_2\frac{T}{e} \right\rfloor$, $m := 0$, $\epsilon_0 := 1$, $\ell_0 = \lceil \frac{2\log(\psi T\epsilon_0^2)}{\epsilon_0} \rceil$ and $N_0 = K * \ell_0$.
Pull each arm once

for $t = K+1, .., T$ **do**

Pull arm i in B_m such that $\min_{i\in B_m}\left\{|\hat{r}_i - \tau| - s_i\right\}$

$t := t+1$

Arm Elimination by Mean Estimation

For each arm $i \in B_m$, remove arm i from B_m if

$$\hat{r}_i + c_i < \tau - c_i$$

For each arm $i \in B_m$, remove arm i from B_m if

$$\hat{r}_i - c_i > \tau + c_i$$

where $c_i = \sqrt{\frac{\rho\log(\psi T\epsilon_m^2)}{2n_i}}$

Arm Elimination by Mean and Variance Estimation

For each arm $i \in B_m$, remove arm i from B_m if

$$\hat{r}_i + s_i < \tau - s_i$$

For each arm $i \in B_m$, remove arm i from B_m if

$$\hat{r}_i - s_i > \tau + s_i$$

where $s_i = \sqrt{\frac{\rho_v\hat{V}_i\log(\psi T\epsilon_m^2)}{4n_i} + \frac{\rho_v\log(\psi T\epsilon_m^2)}{4n_i}}$

if $t \geq N_m$ and $m \leq M$ **then**

Reset Parameters

$\epsilon_{m+1} := \frac{\epsilon_m}{2}$

$B_{m+1} := B_m$

$\ell_{m+1} = \left\lceil \frac{2\log(\psi T\epsilon_{m+1}^2)}{\epsilon_{m+1}} \right\rceil$

$N_{m+1} := t + |B_{m+1}|\ell_{m+1}$

$m := m+1$

end if

end for

Output $\hat{S}_\tau = \{i : \hat{r}_i \geq \tau\}$.

all the arms from A. We divide the entire budget T into rounds/phases as like UCB-Improved, SAR and CSAR. The choice of M comes from UCB-Improved which necessarily entails that the $\epsilon_m \geq \sqrt{\frac{e}{T}}$. So, M is the total number of rounds and is the same as UCB-Improved. After the end of each such round m we eliminate arm(s) from active set B_m and update parameters.

As suggested by [Liu and Tsuruoka, 2016] to make AugUCB an anytime algorithm and to overcome too much early exploration, we no longer pull all the arms equal number of times in each round but pull the arm that minimizes,

$$\min_{i\in B_m}\left\{|\hat{r}_i - \tau| - \sqrt{\frac{\rho_v\hat{V}_i\log(\psi T\epsilon_m^2)}{4n_i} + \frac{\rho_v\log(\psi T\epsilon_m^2)}{4n_i}}\right\}$$

in the active set B_m. This condition makes it possible to pull the arms closer to the threshold τ and with suitable choice of ρ, ρ_v and ψ we can fine tune the exploration. The exact choices for these parameters are derived in Corollary 7.1.1. This is a strategy used by APT. Also, unlike SAR or CSAR, we do not have explicit accept or reject set rather the arm elimination conditions simply removes arm(s) if the algorithm is sufficiently sure that the mean of the arms are very high or very low about the threshold based on mean or variance estimation. This although is a tactic similar to SAR or CSAR, but here at any round, an arbitrary number of arms can be accepted or rejected thereby improving upon SAR and CSAR which accepts/rejects one arm in every round. At the end of the budget T the algorithm outputs all the arms whose estimated average payoff $\hat{r}_i$ is above the threshold τ thereby making this an anytime algorithm whereby we need not finish every round. The arm elimination conditions helps in re-allocating the remaining budget/pulls among the surviving arms.

7 Main Results

7.1 Problem Complexity

We define problem complexity as,

$$H_1 = \sum_{i=1}^{K}\frac{1}{\Delta_i^2}\,,\; H_2 = \max_{i\in A}\frac{i}{\Delta_i^2}, \text{ where } \Delta_i = |r_i - \tau|$$

This is same as the problem complexity defined in [Locatelli *et al.*, 2016] for the thresholding bandit problem and is similar to the problem complexity defined in [Audibert and Bubeck, 2010] for single best arm identification. Also we know that,

$$H_2 \leq H_1 \leq \log(2K)H_2$$

7.2 Theorem 1

Theorem 7.1. *For every $0 < \eta < 1$ and $\gamma > 1$, there exists time t such that for all $T > t$ the simple regret of AugUCB is upper bounded by,*

$$SR_{AugUCB} \leq \sum_{i=1}^{K}\Delta_i\Big\{\exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{16\rho H_2} + \log\big(16\gamma C_1\log_2\frac{T}{e}\big)\Big) + \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) - \frac{c_0\sqrt{T}}{16\rho_v H_2} + \log\big(32\gamma C_2\log_2\frac{T}{e}\big)\Big)\Big\}$$

with probability at least $1-\eta$, where $c_0 > 0$ is a constant and $C_1 = \frac{K\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2})}{T\Delta_i^2}$ and $C_2 = \frac{K\rho_v\log(\psi T\frac{\Delta_i^4}{16\rho_v^2})}{T\Delta_i^2}$.

Proof. According to the algorithm, the number of rounds is $m = \{0, 1, 2, ..M\}$ where $M = \left\lfloor \frac{1}{2}\log_2 \frac{T}{e} \right\rfloor$. So, $\epsilon_m \geq 2^{-M} \geq \sqrt{\frac{e}{T}}$. Also each round m consists of $|B_m|\ell_m$ timesteps where $\ell_m = \frac{\log(\psi T \epsilon_m^2)}{\epsilon_m}$ and B_m is the set of all surviving arms.

Let $c_i = \sqrt{\frac{\rho \log(\psi T \epsilon_m^2)}{2n_i}}$ denote the confidence interval, where n_i is the number of times an arm i is pulled. Let $A^{'} = \{i \in A | \Delta_i \geq b\}$, for $b \geq \sqrt{\frac{e}{T}}$. Let m_i be the minimum round that an arm i gets eliminated such that $m_i = min\{m | \sqrt{\rho\epsilon_m} < \frac{\Delta_i}{2}\}$.

Let $s_i = \sqrt{\frac{\rho_v \hat{V}_i \log(\psi T \epsilon_g^2)}{4n_i} + \frac{\rho_v \log(\psi T \epsilon_g^2)}{4n_i}}$. Let g_i be the minimum round that an arm i gets eliminated such that $g_i = min\{g | \sqrt{\rho_v \epsilon_g} < \frac{\Delta_i}{2}\}$.

Some arm i is not eliminated on or before round $\max\{m_i, g_i\}$

For any arm i, if it is eliminated from active set B_{m_i} then the below two events have to come true,

$$\hat{r}_i + c_i < \tau - c_i, \tag{1}$$
$$\hat{r}_i - c_i > \tau + c_i, \tag{2}$$

For 1 we can see that it eliminates arms that have performed poorly and removes them them from B_{m_i}. Similarly, 2 eliminates arms from B_m that have performed very well compared to threshold τ.

Each round consist of $|B_{m_i}|\ell_{m_i}$ timesteps. In the m_i-th round an arm i can be pulled no more than ℓ_{m_i} times. So when $n_i = \ell_{m_i}$, putting the value of $\ell_{m_i} = \frac{2\log(\psi T \epsilon_{m_i}^2)}{\epsilon_{m_i}}$ in c_i we get,

$$\begin{aligned} c_i &= \sqrt{\frac{\rho\epsilon_{m_i} \log(\psi T \epsilon_{m_i}^2)}{2n_i}} \\ &= \sqrt{\frac{\rho\epsilon_{m_i} \log(\psi T \epsilon_{m_i}^2)}{2 * 2\log(\psi T \epsilon_{m_i}^2)}} \\ &= \frac{\sqrt{\rho\epsilon_{m_i}}}{2} \\ &\leq \sqrt{\rho\epsilon_{m_i+1}} < \frac{\Delta_i}{4}, \text{as } \rho \in (0, 1]. \end{aligned}$$

Again, for $i \in A^{'}$ for 1 elimination condition,

$$\begin{aligned} \hat{r}_i + c_i &\leq r_i + 2c_i = r_i + 4c_i - 2c_i \\ &< r_i + \Delta_i - 2c_i = \tau - 2c_i \leq \tau - c_i \end{aligned}$$

Also, for $i \in A^{'}$ for 2 elimination condition,

$$\begin{aligned} \hat{r}_i - c_i &\geq r_i - 2c_i = r_i - 4c_i + 2c_i \\ &> r_i - \Delta_i + 2c_i \geq \tau + 2c_i \geq \tau + c_i \end{aligned}$$

Now, arm elimination condition is being checked at every timestep, in the m_i-th round as soon as $n_i = \ell_{m_i}$, arm i gets eliminated. Applying Chernoff-Hoeffding bound and considering independence of complementary of the event in 1,

$$\begin{aligned} \mathbb{P}\{\hat{r}_i \geq r_i + 2c_i\} &\leq \exp(-4c_i^2 n_i) \\ &\leq \exp(-8 * \frac{\rho \log(\psi T \epsilon_{m_i}^2)}{2n_i} * n_i) \\ &\leq \exp\big(-4\rho \log(\psi T \epsilon_{m_i}^2)\big) \end{aligned}$$

Similarly, $\mathbb{P}\{\hat{r}_i \leq r_i - 2c_i\} \leq \exp\big(-4\rho \log(\psi T \epsilon_{m_i}^2)\big)$

Summing, the two up, the probability that an arm i is not eliminated on or before m_i-th round based on the 1 and 2 elimination condition is $\left(2\exp\big(-4\rho \log(\psi T \epsilon_{m_i}^2)\big)\right)$.

Again for any arm i, if it is eliminated from active set B_{g_i} then the below two events have to come true,

$$\hat{r}_i + s_i < \tau - s_i, \tag{3}$$
$$\hat{r}_i - s_i > \tau + s_i, \tag{4}$$

For 3 we can see that it eliminates arms that have performed poorly and removes them them from B_{g_i}. Similarly, 4 eliminates arms from B_{g_i} that have performed very well compared to threshold τ.

In the g_i-th round an arm i can be pulled no more than ℓ_{g_i} times. So when $n_i = \ell_{g_i}$, putting the value of $\ell_{g_i} = \frac{2\log(\psi T \epsilon_{g_i}^2)}{\epsilon_{g_i}}$ in s_i we get,

$$\begin{aligned} s_i &= \sqrt{\frac{\rho_v \hat{V}_i \epsilon_{g_i} \log(\psi T \epsilon_{g_i}^2)}{4n_i} + \frac{\rho_v \log(\psi T \epsilon_{g_i}^2)}{4n_i}} \\ &\leq \sqrt{\frac{\rho_v \epsilon_{g_i} \log(\psi T \epsilon_{g_i}^2)}{4 * 2\log(\psi T \epsilon_{g_i}^2)} + \frac{\rho_v \epsilon_{g_i} \log(\psi T \epsilon_{g_i}^2)}{4 * 2\log(\psi T \epsilon_{g_i}^2)}}, \text{as } \hat{V}_i \in [0, 1]. \\ &\leq \sqrt{\frac{\rho_v \epsilon_{g_i}}{8} + \frac{\rho_v \epsilon_{g_i}}{8}} \leq \frac{\sqrt{\rho_v \epsilon_{g_i}}}{2} \\ &\leq \sqrt{\rho_v \epsilon_{g_i+1}} < \frac{\Delta_i}{4}, \text{as } \rho_v \in (0, 1]. \end{aligned}$$

Again, for $i \in A^{'}$ for 3 elimination condition,

$$\begin{aligned} \hat{r}_i + s_i &\leq r_i + 2s_i = r_i + 4s_i - 2s_i \\ &< r_i + \Delta_i - 2s_i = \tau - 2s_i \leq \tau - s_i \end{aligned}$$

Also, for $i \in A^{'}$ for 4 elimination condition,

$$\begin{aligned} \hat{r}_i - s_i &\geq r_i - 2s_i = r_i - 4s_i + 2s_i \\ &> r_i - \Delta_i + 2s_i \geq \tau + 2s_i \geq \tau + s_i \end{aligned}$$

Since, arm elimination condition is being checked at every timestep, in the g_i-th round as soon as $n_i = \ell_{g_i}$, arm i gets eliminated. Applying Bernstein inequality and considering independence of complementary of the event in 3,

$$\mathbb{P}\{\hat{r}_i \geq r_i + 2s_i\} \tag{5}$$
$$\leq \mathbb{P}\left\{\hat{r}_i \geq r_i + (2\sqrt{\frac{\rho_v \hat{V}_i \log(\psi T \epsilon_{g_i}^2) + \rho_v \log(\psi T \epsilon_{g_i}^2)}{4n_i}})\right\} \tag{6}$$
$$\leq \mathbb{P}\left\{\hat{r}_i \geq r_i + (2\sqrt{\frac{\rho_v [\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1] \log(\psi T \epsilon_{g_i}^2)}{4n_i}})\right\} \tag{7}$$

$$+ \mathbb{P}\Bigg\{\hat{V}_i \geq \sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}\Bigg\} \tag{8}$$

Now, we know that in the g_i-th round,

$$2\sqrt{\frac{\rho_v[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}]\log(\psi T\epsilon_{g_i}^2)}{4n_i} + \frac{\rho_v \log(\psi T\epsilon_{g_i}^2)}{4n_i}}$$
$$\leq 2\sqrt{\frac{\rho_v[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}]\log(\psi T\epsilon_{g_i}^2)}{\frac{8\log(\psi T\epsilon_{g_i}^2)}{\epsilon_{g_i}}} + \frac{\rho_v \log(\psi T\epsilon_{g_i}^2)}{\frac{8\log(\psi T\epsilon_{g_i}^2)}{\epsilon_{g_i}}}}$$
$$\leq \frac{\sqrt{\rho_v\epsilon_{g_i}[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1]}}{2} \leq \sqrt{\rho_v\epsilon_{g_i}}$$

For the term in 7, by applying Bernstein inequality, we can write as,

$$\mathbb{P}\Bigg\{\hat{r}_i \geq r_i + \Bigg(2\sqrt{\frac{\rho_v[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1]\log(\psi T\epsilon_{g_i}^2)}{4n_i}}\Bigg)\Bigg\}$$
$$\leq \exp\Bigg(-\frac{\Bigg(2\sqrt{\frac{\rho_v[\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}]\log(\psi T\epsilon_{g_i}^2)}{4n_i}} + \frac{\rho_v\log(\psi T\epsilon_{g_i}^2)}{4n_i}\Bigg)^2 n_i}{2\sigma_i^2 + \frac{4}{3}\sqrt{\frac{\rho_v[\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}]\log(\psi T\epsilon_{g_i}^2)}{4n_i}} + \frac{\rho_v\log(\psi T\epsilon_{g_i}^2)}{4n_i}}\Bigg)$$
$$\leq \exp\Bigg(-\frac{\Big(\rho_v[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1]\log(\psi T\epsilon_{g_i}^2)\Big)}{2\sigma_i^2 + \frac{2}{3}\sqrt{\rho_v\epsilon_{g_i}}}\Bigg)$$
$$\leq \exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\Bigg)\log(\psi T\epsilon_{g_i}^2)\Bigg)$$

For the term in 8, by applying Bernstein inequality, we can write as,

$$\mathbb{P}\Bigg\{\hat{V}_i \geq \sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}\Bigg\}$$
$$\leq \mathbb{P}\Bigg\{\frac{1}{n_i}\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2 - (\hat{r}_i - r_i)^2 \geq \sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}\Bigg\}$$
$$\leq \mathbb{P}\Bigg\{\frac{\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}\Bigg\}$$
$$\leq \mathbb{P}\Bigg\{\frac{\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 +$$
$$\Bigg(2\sqrt{\frac{\rho_v[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}]\log(\psi T\epsilon_{g_i}^2)}{4n_i} + \frac{\rho_v\log(\psi T\epsilon_{g_i}^2)}{4n_i}}\Bigg)\Bigg\}$$
$$\leq \exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\Bigg)\log(\psi T\epsilon_{g_i}^2)\Bigg)$$

Similarly, the condition for the complementary event for the elimination case 4 holds such that $\mathbb{P}\{\hat{r}_i \leq r_i - 2s_i\} \leq 2\exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\Bigg)\log(\psi T\epsilon_{g_i}^2)\Bigg)$.

Summing everything up, the probability that an arm i is not eliminated on or before g_i-th round based on the 3 and 4 elimination condition is $4\exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\Bigg)\log(\psi T\epsilon_{g_i}^2)\Bigg)$.

1. *Fact 1:* From above we know that the probability of elimination of a sub-optimal arm in the $\max\{m_i, g_i\}$-th round being not eliminated is bounded above by $P_{m_i} \leq 2\exp\Bigg(-4\rho\log(\psi T\epsilon_{m_i}^2)\Bigg) + 4\exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\Bigg)\log(\psi T\epsilon_{g_i}^2)\Bigg)$.
2. *Fact 2:* From [Tolpin and Shimony, 2012] we know that, for every $0 < \eta < 1$ and $\gamma > 1$, there exists t such that for all $T > t$ the probability of a sub-optimal arm i being sampled in the m_i-th round is bounded by $Q_{m_i} \leq 2\gamma\exp(-c_{m_i}\frac{\sqrt{T}}{2})$, where $c_{m_i} = \frac{c_0}{2^{m_i}}$.

We start with an upper bound on the number of plays $\delta_{\max\{m_i,g_i\}}$ in the $\max\{m_i, g_i\}$-th round divided by the total number of plays T. We know from Fact 1 that the total number of arms surviving in the $\max\{m_i, g_i\}$-th arm is,

$$|B_{\max\{m_i,g_i\}}| = 2K\exp\Bigg(-4\rho\log(\psi T\epsilon_{m_i}^2)\Bigg)$$
$$+ 4K\exp\Bigg(-\frac{3\rho_v}{2}(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}})\log(\psi T\epsilon_{g_i}^2)\Bigg)$$

Again for AugUCB, we know that the number of pulls allocated for each surviving arm i in the m_i-th round is $\ell_{m_i} = \frac{2\log(\psi T\epsilon_{m_i}^2)}{\epsilon_{m_i}}$ or for the g_i-th round is $\ell_{g_i} = \frac{2\log(\psi T\epsilon_{g_i}^2)}{\epsilon_{g_i}}$. Therefore, the proportion of plays $\delta_{\max\{m_i,g_i\}}$ in the $\max\{m_i, g_i\}$-th round can be written as,

$$\delta_{\max\{m_i,g_i\}} = \frac{(|B_{m_i}|.\ell_{m_i})}{T} + \frac{(|B_{g_i}|.\ell_{g_i})}{T}$$
$$\leq \frac{2K}{T}\exp\Bigg(-4\rho\log(\psi T\epsilon_{m_i}^2)\Bigg).\frac{2\log(\psi T\epsilon_{m_i}^2)}{\epsilon_{m_i}}$$
$$+ \frac{4K}{T}\exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\Bigg)\log(\psi T\epsilon_{g_i}^2)\Bigg).\frac{2\log(\psi T\epsilon_{g_i}^2)}{\epsilon_{g_i}}$$
$$\leq \frac{4K\log(\psi T\epsilon_{m_i}^2)}{T\epsilon_{m_i}}\exp\Bigg(-4\rho\log(\psi T\epsilon_{m_i}^2)\Bigg)$$
$$+ \frac{8K\log(\psi T\epsilon_{g_i}^2)}{T\epsilon_{g_i}}\exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\Bigg)\log(\psi T\epsilon_{g_i}^2)\Bigg)$$

Now, in the $\max\{m_i, g_i\}$-th round $\sqrt{\rho\epsilon_{m_i}} \leq \frac{\Delta_i}{2}$ or $\sqrt{\rho_v\epsilon_{g_i}} \leq \frac{\Delta_i}{2}$. Hence,

$$\delta_{\max\{m_i,g_i\}} \leq \frac{4K\log(\psi T\frac{\Delta_i^4}{16\rho^2})}{T\frac{\Delta_i^2}{4\rho}}\exp\Bigg(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2})\Bigg)$$
$$+ \frac{8K\log(\psi T\frac{\Delta_i^4}{16\rho_v^2})}{T\frac{\Delta_i^2}{4\rho_v}}\exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{\sigma_i^2 + \frac{\Delta_i}{2} + 1}{3\sigma_i^2 + \frac{\Delta_i}{2}}\Bigg)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2})\Bigg)$$
$$\leq 16C_1\exp\Bigg(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2})\Bigg)$$
$$+ 32C_2\exp\Bigg(-\frac{3\rho_v}{2}\Bigg(\frac{2\sigma_i^2 + \Delta_i + 2}{6\sigma_i^2 + \Delta_i}\Bigg)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2})\Bigg)$$

where $C_1 = \frac{K\rho \log(\psi T \frac{\Delta_i^4}{16\rho^2})}{T\Delta_i^2}$ and $C_2 = \frac{K\rho_v \log(\psi T \frac{\Delta_i^4}{16\rho_v^2})}{T\Delta_i^2}$

Now, applying the bound from Fact 2, we can show that for all rounds $m = 0, 1, 2, ..., M$ the probability of the sub-optimal arm i being pulled is bounded above by,

$$
\begin{aligned}
&P_i = \sum_{m=0}^{M} \delta_{m_i}.Q_{m_i} + \sum_{g=0}^{M} \delta_{g_i}.Q_{g_i} \\
&\leq \sum_{m=0}^{M} \Bigg\{ 16C_1 \exp\Big(-4\rho \log(\psi T \frac{\Delta_i^4}{16\rho^2})\Big).2\gamma \exp(-\frac{c_0\sqrt{T}}{2^{m_i}.4}) \\
&+ 32C_2 \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2})\Big). \\
&2\gamma \exp(-\frac{c_0\sqrt{T}}{2^{g_i}.4})\Bigg\} \\
&\leq M\gamma\Bigg\{ 32C_1 \exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{\frac{4\rho}{\Delta_i^2}.4}\Big) \\
&+ 64C_2 \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) \\
&- \frac{c_0\sqrt{T}}{\frac{4\rho_v}{\Delta_i^2}.4}\Big)\Bigg\}, \text{as } \frac{1}{2^{m_i}} = \epsilon_{m_i} \text{ or } \frac{1}{2^{g_i}} = \epsilon_{g_i} \\
&\leq \gamma \log_2 \frac{T}{e}\Bigg\{\Big(16C_1 \exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{16\rho\Delta_i^{-2}}\Big) \\
&+ 32C_2 \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) \\
&- \frac{c_0\sqrt{T}}{16\rho_v\Delta_i^{-2}}\Big)\Bigg\} \text{ for } M = \lfloor \frac{1}{2}\log_2\frac{T}{e}\rfloor \\
&\leq \gamma \log_2 \frac{T}{e}\Bigg\{ 16C_1 \exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{16\rho i \max_i \Delta_i^{-2}}\Big) \\
&+ 32C_2 \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) \\
&- \frac{c_0\sqrt{T}}{16\rho_v i \max_i \Delta_i^{-2}}\Big)\Bigg\} \\
&\leq \gamma \log_2 \frac{T}{e}\Bigg\{ 16C_1 \exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{16\rho H_2}\Big) \\
&+ 32C_2 \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) \\
&- \frac{c_0\sqrt{T}}{16\rho_v H_2}\Big)\Bigg\} \\
&\leq \exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{16\rho H_2} + \log\big(16\gamma C_1 \log_2\frac{T}{e}\big)\Big) \\
&+ \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) \\
&- \frac{c_0\sqrt{T}}{16\rho_v H_2} + \log\big(32\gamma C_2 \log_2\frac{T}{e}\big)\Big)
\end{aligned}
$$

Therefore we can say that with probability $1 - P_i$, all arms i above $\frac{\Delta_i}{2}$ are accepted and all arms i below $\frac{\Delta_i}{2}$ are rejected.

Hence, the simple regret of AugUCB is upper bounded by,

$$
\begin{aligned}
&SR_{AugUCB} \leq \sum_{i=1}^{K} \Delta_i.P_i \\
&\leq \sum_{i=1}^{K} \Delta_i \Bigg\{ \exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{16\rho H_2} \\
&+ \log\big(16\gamma C_1 \log_2\frac{T}{e}\big)\Big) + \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) \\
&- \frac{c_0\sqrt{T}}{16\rho_v H_2} + \log\big(32\gamma C_2 \log_2\frac{T}{e}\big)\Big)\Bigg\}
\end{aligned}
$$

□

Next we specialize the result of Theorem 7.1 in Corollary 7.1.1.

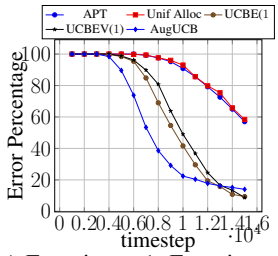
7.3 Corollary 2

Corollary 7.1.1. *For $c_0 = \sqrt{T}$, $\psi = \frac{T}{\log(K)}$, $\rho = \frac{1}{8}$ and $\rho_v = \frac{2}{3}$, the simple regret of AugUCB is given by,*

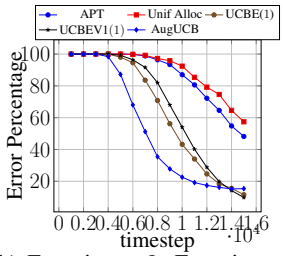
$$
\begin{aligned}
SR_{AugUCB} &\leq \sum_{i=1}^{K} \Delta_i \Bigg\{ \exp\Big(-\log(2T\frac{\Delta_i^2}{\sqrt{\log K}}) - \frac{T}{2H_2} \\
&+ \log\big(\frac{4\gamma K \log(2T\frac{\Delta_i^2}{\sqrt{\log K}})}{T\Delta_i^2}\log_2\frac{T}{e}\big)\Big) \\
&+ \exp\Big(-\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(3T\frac{\Delta_i^2}{8\sqrt{\log K}}) - \frac{3T}{32H_2} \\
&+ \log\big(\frac{64\gamma K \log(3T\frac{\Delta_i^2}{8\sqrt{\log K}})}{3T\Delta_i^2}\log_2\frac{T}{e}\big)\Big)\Bigg\}
\end{aligned}
$$

Proof. Putting $c_0 = \sqrt{T}$, $\psi = \frac{T}{\log(K)}$, $\rho = \frac{1}{8}$ and $\rho_v = \frac{2}{3}$ in the result obtained in Theorem 7.1, we get

$$
\begin{aligned}
&SR_{AugUCB} \leq \sum_{i=1}^{K} \Delta_i \Bigg\{ \exp\Big(-4\rho\log(\psi T\frac{\Delta_i^4}{16\rho^2}) - \frac{c_0\sqrt{T}}{16\rho H_2} \\
&+ \log\big(16\gamma C_1 \log_2\frac{T}{e}\big)\Big) + \exp\Big(-\frac{3\rho_v}{2}\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(\psi T\frac{\Delta_i^4}{16\rho_v^2}) \\
&- \frac{c_0\sqrt{T}}{16\rho_v H_2} + \log\big(32\gamma C_2 \log_2\frac{T}{e}\big)\Big)\Bigg\} \\
&\leq \sum_{i=1}^{K} \Delta_i \Bigg\{ \exp\Big(-\frac{1}{2}\log(T^2\frac{4\Delta_i^4}{\log K}) - \frac{T}{2H_2} \\
&+ \log\big(\frac{2\gamma K \log(T^2\frac{4\Delta_i^4}{\log K})}{T\Delta_i^2}\log_2\frac{T}{e}\big)\Big) \\
&+ \exp\Big(-\Big(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\Big)\log(T^2\frac{\Delta_i^4}{16.\frac{4}{9}\log K}) - \frac{c_0\sqrt{T}}{16.\frac{2}{3}H_2} \\
&+ \log\big(\frac{32\gamma\rho_v K \log(T^2\frac{\Delta_i^4}{16.\frac{2}{9}\log K})}{T\Delta_i^2}\log_2\frac{T}{e}\big)\Big)\Bigg\} \\
&\leq \sum_{i=1}^{K} \Delta_i \Bigg\{ \exp\Big(-\log(2T\frac{\Delta_i^2}{\sqrt{\log K}}) - \frac{T}{2H_2} \\
&+ \log\big(\frac{4\gamma K \log(2T\frac{\Delta_i^2}{\sqrt{\log K}})}{T\Delta_i^2}\log_2\frac{T}{e}\big)\Big)
\end{aligned}
$$



(a) Experiment 1: Experiment with Arithmetic Progression



(b) Experiment 2: Experiment with 4 Group Setting

Figure 1: Experiments with thresholding bandit

$$+\exp\left(-\left(\frac{2\sigma_i^2+\Delta_i+2}{6\sigma_i^2+\Delta_i}\right)\log(3T\frac{\Delta_i^2}{8\sqrt{\log K}})-\frac{3T}{32H_2}\right.$$
$$\left.+\log\left(\frac{64\gamma K\log(3T\frac{\Delta_i^2}{8\sqrt{\log K}})}{3T\Delta_i^2}\log_2\frac{T}{e}\right)\right)\Bigg\}$$

□

8 Experimental Run:

In this section we compare the empirical performance of AugUCB against APT, Uniform Allocation, UCBE and UCBEV algorithm. The threshold τ is set at 0.5 for all experiments. Each algorithm is run independently 500 times for 15000 timesteps and the output set of arms suggested by the algorithms at every timestep is recorded. The output is considered erroneous if the correct set of arms is not $i = \{6, 7, 8, 9, 10\}$ (true for all the experiments). The error percentage over 500 runs is plotted against 15000 timesteps. For the uniform allocation algorithm, for each $t = 1, 2, .., T$ the arms are sampled uniformly. For UCBE algorithm ([Audibert *et al.*, 2009]) which was built for single best arm identification, we modify it according to [Locatelli *et al.*, 2016] to suit the goal of finding arms above the threshold τ. So the exploration parameter a in UCBE is set to $a_i = \frac{T-K}{H_1}$. Then for each timestep $t = 1, 2, .., T$ we pull the arm that maximizes $\{|\hat{r}_i - \tau| - \sqrt{\frac{a_i}{n_i}}\}$, where n_i is the number of times the arm i is pulled till $t-1$ timestep. Also, APT is run with $\epsilon = 0$, which denotes the precision with which the algorithm suggests the best set of arms. So when ϵ is set to 0 APT has to suggest the exact set of arms above the threshold. For AugUCB we take $\psi = \frac{T}{\log K}$, $\rho = \frac{1}{8}$ and $\rho_v = \frac{2}{3}$ as in Corollary 7.1.1.

The first experiment is conducted on a testbed of 100 arms involving Gaussian reward distribution with expected rewards of the arms $r_{1:4} = 0.2 + (0:3) * 0.05$, $r_5 = 0.45$, $r_6 = 0.55$, $r_{7:10} = 0.65 + (0:3) * 0.05$ and $r_{11:100}$=0.4. The means of first 10 arms are set as arithmetic progression. Variance is set as $\sigma^2_{i=1:6 \text{ and } 11:100} = 0.5$ and $\sigma_{i=7:10} = 0.6$. The means in the testbed are chosen in such a way that any algorithm has to spend a significant amount of budget to explore all the arms and variance is chosen in such a way that the arms above τ have high variance. In this experiment we see that AugUCB performs better than all the other algorithms mentioned. Only UCBE(1) and UCBEV(1) beats AugUCB and that is because they have access to the exact problem complexity. The result is shown in Figure 1(a).

The second experiment is conducted on a testbed of 100 arms with the means divided into 4 groups. Again the testbed involves Gaussian reward distributions with expected rewards of the arms as $r_{1:3} = 0.1$, $r_{4:7} = \{0.35, 0.45, 0.55, 0.65\}$, $r_{8:10} = 0.9$ and $r_{11:100} = 0.4$. Also $\sigma^2_{i=1:7 \text{ and } 11:100} = 0.5$ and $\sigma^2_{i=8:10} = 0.6$. AugUCB, APT, Uniform Allocation, UCBE and UCBEV with the same settings as experiment 1 are run on this testbed. The result is shown in Figure 1(b). Here, also we see that AugUCB beats APT.

9 Conclusion and Future work

To be written.

References

[Audibert and Bubeck, 2009] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

[Audibert and Bubeck, 2010] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

[Audibert *et al.*, 2009] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[Auer and Ortner, 2010] Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

[Auer *et al.*, 2002a] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[Auer *et al.*, 2002b] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[Bubeck and Cesa-Bianchi, 2012] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[Bubeck *et al.*, 2011] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[Bubeck *et al.*, 2013] Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *ICML (1)*, pages 258–265, 2013.

[Chen *et al.*, 2014] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014.

[Even-Dar *et al.*, 2006] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7:1079–1105, 2006.

[Gabillon *et al.*, 2011] Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2011.

[Gabillon *et al.*, 2012] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

[Kalyanakrishnan *et al.*, 2012] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 655–662, 2012.

[Lai and Robbins, 1985] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[Liu and Tsuruoka, 2016] Yun-Ching Liu and Yoshimasa Tsuruoka. Modification of improved upper confidence bounds for regulating exploration in monte-carlo tree search. *Theoretical Computer Science*, 2016.

[Locatelli *et al.*, 2016] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. *arXiv preprint arXiv:1605.08671*, 2016.

[Robbins, 1952] Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1952.

[Thompson, 1933] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

[Tolpin and Shimony, 2012] David Tolpin and Solomon Eyal Shimony. Mcts based on simple regret. In *AAAI*, 2012.