Thresholding Bandits with Augmented UCB

Author names withheld

Abstract

In this paper we introduce the algorithm Augmented UCB for the combinatorial fixed budget pure exploration stochastic multi-armed bandit setup called thresholding bandit problem. Our algorithm is an anytime action elimination variance-aware algorithm and is the first of its kind which employs multiple action elimination conditions based on estimated mean and variance of the arms. Theoretically, our algorithm provides a weaker guarantee than APT[Locatelli *et al.*, 2016] and CSAR[Chen *et al.*, 2014], but empirically it performs much better since it is a variance-aware algorithm. In the testcases involving large set of arms our algorithm has consistently performed much better than the state-of-the art APT and CSAR algorithms.

1 Introduction

In this paper we study a specific combinatorial pure-exploration problem called the thresholding bandit problem (TBP) in the context of stochastic multi-armed bandit (MAB) setting. MAB problems are instances of the classic sequential decision-making scenario. Specifically, a MAB problem comprises a learner and a collection of actions (or arms), denoted $\mathcal{A}$; subsequent plays (or pulls) of an arm $i \in \mathcal{A}$ yields independent and identically distributed (i.i.d.) reward samples from a distribution (corresponding to arm i), whose expectation is denoted by r_i. The learner's objective is to identify an arm corresponding to the maximum expected reward, denoted r^*. Thus, at each time-step the learner is faced with the *exploration vs. exploitation dilemma*, whereby it can pull an arm which has yielded the highest mean reward (denoted $\hat{r}_i$) thus far (*exploitation*) or continue to explore other arms with the prospect of finding a better arm whose performance is yet not observed sufficiently (*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 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. 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.

1.1 Motivation

The 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). From areas ranging from Anomaly Detection and Classification ([Locatelli *et al.*, 2016]) to industrial application where the learner has to keep all those workers active whose productivity is above a particular threshold τ, or allocating channels whose quality is above a threshold for Mobile Communications ([Audibert and Bubeck, 2010]) or in crowd-sourcing while hiring workers the TBP problem can be employed.

1.2 Related Work

A significant amount of work has been done on the stochastic MAB 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]. Starting from the early work of [Thompson, 1933], [Robbins, 1952] to [Lai and Robbins, 1985] which gives us an asymptotic lower bound on the cumulative regret we come to the UCB1 algorithm in [Auer *et al.*, 2002]. Subsequent works such as [Audibert and Bubeck, 2009] and [Auer and Ortner, 2010] have shown better upper bounds on the cumulative regret. In [Auer and Ortner, 2010] they propose the UCB-Improved algorithm which is round-based algorithm[1]. Of special mention is the [Audibert *et al.*, 2009] where they introduce variance-aware algorithm UCB-V and show that algorithms that take into account variance estimation along with mean estimation tends to perform better than algorithms than solely focuses on mean estimation such as UCB1.

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). We study this setting in our paper. In [Audibert and Bubeck, 2010] the authors come up with the algorithm UCBE and Successive Reject(SR) with simple regret guarantees to find the single best arm. 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 a 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. A similar combinatorial setup was also explored in [Chen *et al.*, 2014] where the authors come up with an algorithm, called Combinatorial Successive Accept Reject (CSAR) which is similar to SAR but with a more general setup.

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] while in the combinatorial setup [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. For a survey of this setup we refer the reader to [Jamieson and Nowak, 2014].

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. 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.

[1] An algorithm is *round-based* if it pulls all the arms equal number of times in each round and then proceeds to eliminate one or more arms that it identifies to be sub-optimal.

1.3 Our Contribution

In this paper we propose the Algorithm AugUCB which is an anytime action elimination algorithm suited for the TBP problem. It combines the approach of UCB-Improved, CCB ([Liu and Tsuruoka, 2016]) and APT algorithm. Our algorithm is a variance-aware algorithm which takes into account the empirical variance of the arms. We also address an open problem raised in [Auer and Ortner, 2010] of coming up with an algorithm that can eliminate arms based on variance. Both CSAR and APT are not variance-aware algorithms. The expected loss of various algorithms is shown in Table 1. The terms H_1, H_2, H_1^σ and H_2^σ signifies problem complexity and are defined in section 4. Theoretically, we can compare the first term (containing H_2) of our expected loss and see that for all $K \geq 3$, $\frac{K\log K}{\log(2K\log K)} > \log 2K > logK \geq \bar{\log} H_2 \geq H_1$ and hence our result is weaker than CSAR and APT. The term containing H_2^σ is comparable to the Gap-EV([Gabillon *et al.*, 2011] or UGapE-V([Gabillon *et al.*, 2012]) algorithm. In both these algorithms we get the term H_1^σ which is less than our definition of $H_2(\sigma)$ and hence our algorithm is weaker with respect to Gap-EV or UGapE-V. But of these algorithm needs the complexity factor H_1^σ as input for optimal performance (which is not a realistic scenario) whereas AugUCB eliminates arms depending on empirical variance estimate of arms.

Table 1: Expected Loss for different bandit algorithms

Algorithm	Upper Bound on Expected Loss
APT	$\exp(-\frac{T}{64H_1} + 2\log((\log(T)+1)K))$
CSAR	$K^2\exp(-\frac{T-K}{72\log(K)H_2})$
AugUCB	$\exp\left(-\frac{T\log(2K\log K)}{16H_2K\log K} + \log\left(K\left(\log_2\frac{T}{e}\right)\right)\right)$ $+$ $\exp\left(-\frac{T\log(\frac{3}{4}K\log K)}{64H_2^\sigma K(\log K)} + 2\log\left(K\left(\log_2\frac{T}{e}\right)\right)\right)$

Empirically we show that for a large action set when the variance of the arms lying above τ are high, our algorithm performs better than all other algorithms, except the algorithm UCBEV which has access to the underlying problem complexity and also is a variance-aware algorithm. Irrespective of this case AugUCB also employs elimination of arms based on mean estimation only and is the first such algorithm which uses elimination by both mean and variance estimation simultaneously. AugUCB requires three input parameters and the exact choices for these parameters are derived in Theorem 4.1. Also, unlike SAR or CSAR, AugUCB does not have explicit accept or reject set rather the arm elimination conditions simply removes arm(s) if it is sufficiently sure that the mean of the arms are very high or very low about the threshold based on mean and variance estimation thereby reallocating the remaining budget among the surviving arms. 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. Also their round lengths are non-adaptive and they pull all the arms equal number of times in each round. The rest of the paper is divided as follows, in section 2 we introduce the notations and the in the next section 3 we present AugUCB. Section 4 contains our main theorem on expected loss, section 5 contains the numerical experiments and we conclude in section 6.

2 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 $\hat{\Delta}_i = |\tau - \hat{r}_i|$. n_i denotes the number of times the arm i has been pulled. ψ denotes the exploration regulatory factor and ρ_μ, ρ_v as arm elimination parameters. $\hat{V}_i = \frac{1}{n_i}\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2$ 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 i. It is assumed that the distribution from which rewards are sampled are identical and independent 1-sub-Gaussian distributions which includes 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 $[0, 1]$.

3 Augmented UCB

In algorithm 1, hence referred to as AugUCB, we have three exploration parameters, ρ_μ, ρ_v which are the arm elimination parameters and ψ which is the exploration regulatory factor. The main approach is based on UCB-Improved with modifications suited for the thresholding bandit problem. The active set B_0 is initialized with all the arms from A. We divide the entire budget T into rounds/phases as like UCB-Improved, CCB, SAR and CSAR. 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, $\min_{i\in B_m}\left\{|\hat{r}_i - \tau| - 2\sqrt{\frac{\rho_v\psi\hat{V}_i\log(T\epsilon_m)}{4n_i} + \frac{\rho_v\psi\log(T\epsilon_m)}{4n_i}}\right\}$ 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_μ, ρ_v and ψ we can fine tune the exploration. Also because of the said condition, like [Liu and Tsuruoka, 2016] we also claim that AugUCB is an anytime algorithm. The choice of exploration factor ψ comes directly from [Audibert and Bubeck, 2010] and [Bubeck *et al.*, 2011] which states that in pure exploration setup, the exploring factor must be linear in T to give us an exponentially small probability of error rather than logarithmic which is suited for minimizing cumulative regret.

Algorithm 1 AugmentedUCB

Input: Time horizon T, exploration parameters ρ_μ, ρ_v and ψ, threshold τ.
Initialization: Set $B_0 := A$, $M = \left\lfloor \frac{1}{2}\log_2 \frac{T}{e} \right\rfloor$, $m := 0$, $\epsilon_0 := 1$, $\ell_0 = \left\lceil \frac{2\psi\log(T\epsilon_0)}{\epsilon_0} \right\rceil$ and $N_0 = K\ell_0$.
Pull each arm once

for $t = K+1, .., T$ **do**
  Pull arm $i \in \arg\min_{j\in B_m}\left\{|\hat{r}_j - \tau| - 2s_j\right\}$
  $t := t + 1$
  Arm Elimination by Mean Estimation
    For each arm $i \in B_m$, remove arm i from B_m if

$$\hat{r}_i + c_i < \tau - c_i \text{ or } \hat{r}_i - c_i > \tau + c_i$$
$$\text{where } c_i = \sqrt{\frac{\rho_\mu\psi\log(T\epsilon_m)}{2n_i}}$$

  Arm Elimination by Mean and Variance Estimation
    For each arm $i \in B_m$, remove arm i from B_m if

$$\hat{r}_i + s_i < \tau - s_i \text{ or } \hat{r}_i - s_i > \tau + s_i$$
$$\text{where } s_i = \sqrt{\frac{\rho_v\psi\hat{V}_i\log(T\epsilon_m)}{4n_i} + \frac{\rho_v\psi\log(T\epsilon_m)}{4n_i}}$$

  if $t \geq N_m$ and $m \leq M$ **then**
    Reset Parameters
      $\epsilon_{m+1} := \frac{\epsilon_m}{2}$
      $B_{m+1} := B_m$
      $\ell_{m+1} := \left\lceil \frac{2\psi\log(T\epsilon_{m+1})}{\epsilon_{m+1}} \right\rceil$
      $N_{m+1} := t + |B_{m+1}|\ell_{m+1}$
      $m := m + 1$
  end if
end for
Output $\hat{S}_\tau = \{i : \hat{r}_i \geq \tau\}$.

4 Main Results

4.1 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,

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

Also, we define H_1^σ ([Gabillon *et al.*, 2011]) and H_2^σ as,

$$H_1^\sigma = \sum_{i=1}^{K}\frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}$$

$$H_2^\sigma = \max_{i \in A} i \frac{6\sigma_i^2 + \Delta_i}{6\Delta_i^2}$$

which also gives us that $H_2^\sigma < H_1^\sigma$.

4.2 Theorem 1

Theorem 4.1. *With $\psi = \frac{T\epsilon_m}{8K\log K}$, $\rho_\mu = \frac{1}{8}$ and $\rho_v = \frac{1}{3}$, the expected loss of the AugUCB algorithm is given by,*

$$\mathbb{E}[\mathcal{L}(T)] \leq \exp\left(-\frac{T\log(2K\log K)}{16H_2K\log K} + \log\left(K\big(\log_2 \frac{T}{e} + 1\big)\right)\right) + \exp\left(-\frac{T\log(\frac{3}{4}K\log K)}{64H_2^\sigma K(\log K)} + 2\log\left(K\big(\log_2 \frac{T}{e} + 1\big)\right)\right).$$

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 $\ell_m = \left\lceil \frac{2\psi\log(T\epsilon_m)}{\epsilon_m} \right\rceil$ and B_m is the set of all surviving arms.

Let $c_i = \sqrt{\frac{\rho_\mu\psi\log(T\epsilon_m)}{2n_i}}$ 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 $b \geq \sqrt{\frac{e}{T}}$. Define $m_i = \min\{m | \sqrt{\rho_\mu\epsilon_m} < \frac{\Delta_i}{2}\}$.

Let $s_i = \sqrt{\frac{\rho_v\psi\hat{V}_i\log(T\epsilon_g)}{4n_i} + \frac{\rho_v\psi\log(T\epsilon_g)}{4n_i}}$ and $g_i = min\{g | \sqrt{\rho_v\epsilon_g} < \frac{\Delta_i}{2}\}$.

Let ξ_1 and ξ_2 be the favorable event such that,

$$\xi_1 = \left\{ \forall i \in A, \forall m = 0, 1, 2, .., M : |\hat{r}_i - r_i| \leq 2c_i \right\}$$

$$\xi_2 = \left\{ \forall i \in A, \forall m = 0, 1, 2, .., M : |\hat{r}_i - r_i| \leq 2s_i \right\}$$

So, ξ_1 and ξ_2 signifies the event till any arm i will get eliminated from B_m.

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 one of the below two events has to occur,

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

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

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

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 $\ell_{m_i} \geq \frac{2\psi\log(T\epsilon_{m_i})}{\epsilon_{m_i}}$ in c_i we get,

$$c_i = \sqrt{\frac{\rho_\mu\psi\epsilon_{m_i}\log(T\epsilon_{m_i})}{2n_i}} \leq \sqrt{\frac{\rho_\mu\psi\epsilon_i\log(T\epsilon_{m_i})}{2 * 2\psi\log(T\epsilon_{m_i})}}$$
$$\leq \frac{\sqrt{\rho_\mu\epsilon_{m_i}}}{2} < \frac{\Delta_i}{4}, \text{ as } \rho_\mu \in (0, 1].$$

Again, for $i \in A'$ for the elimination condition in (1),

$$\hat{r}_i \leq r_i + 2c_i = r_i + 4c_i - 2c_i$$
$$< r_i + \Delta_i - 2c_i = \tau - 2c_i.$$

Similarly, for $i \in A'$ for the elimination condition in (2),

$$\hat{r}_i \geq r_i - 2c_i = r_i - 4c_i + 2c_i$$
$$> r_i - \Delta_i + 2c_i = \tau + 2c_i.$$

Applying Chernoff-Hoeffding bound and considering independence of complementary of the event in (1),

$$\mathbb{P}\{\hat{r}_i > r_i + 2c_i\} \leq \exp(-4c_i^2 n_i)$$
$$\leq \exp(-8 * \frac{\rho_\mu\psi\log(T\epsilon_{m_i})}{2n_i} * n_i)$$
$$\leq \exp\big(-4\rho_\mu\psi\log(T\epsilon_{m_i})\big)$$
$$\leq \exp\left(-\rho_\mu\frac{T\epsilon_{m_i}}{2K\log K}\log(T\epsilon_{m_i})\right),$$

putting the value of $\psi = \frac{T\epsilon_{m_i}}{8K\log K}$

Similarly for the condition in (2), $\mathbb{P}\{\hat{r}_i < r_i - 2c_i\} \leq \exp\left(-\frac{T\rho_\mu\epsilon_{m_i}}{2K\log K}\log(T\epsilon_{m_i})\right)$.

Summing the above two expressions, the probability that arm i is not eliminated on or before m_i-th is $\left(2\exp\left(-\frac{4T\rho_\mu\epsilon_{m_i}}{8K\log K}\log(T\epsilon_{m_i})\right)\right)$.

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

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

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

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 $\ell_{g_i} \geq \frac{2\psi\log(T\epsilon_{g_i})}{\epsilon_{g_i}}$ in s_i we get,

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

Again, for $i \in A'$ for the elimination condition in (3),

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

Also, for $i \in A'$ for the elimination condition in (4),

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

Applying Bernstein inequality and considering independence of complementary of the event in (3),

$$\mathbb{P}\{\hat{r}_i > r_i + 2s_i\} \tag{5}$$

$$\leq \mathbb{P}\bigg\{\hat{r}_i > r_i + (2\sqrt{\frac{\rho_v\psi\hat{V}_i\log(T\epsilon_{g_i}) + \rho_v\psi\log(T\epsilon_{g_i})}{4n_i}})\bigg\} \tag{6}$$

$$\leq \mathbb{P}\bigg\{\hat{r}_i > r_i + (2\sqrt{\frac{\rho_v\psi[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1]\log(T\epsilon_{g_i})}{4n_i}})\bigg\} \tag{7}$$

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

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

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

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

$$\mathbb{P}\bigg\{\hat{r}_i > r_i + \bigg(2\sqrt{\frac{\rho_v\psi[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1]\log(T\epsilon_{g_i})}{4n_i}}\bigg)\bigg\}$$
$$\leq \exp\Bigg(-\frac{\bigg(2\sqrt{\frac{\rho_v\psi[\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}]\log(T\epsilon_{g_i})}{4n_i} + \frac{\rho_v\psi\log(T\epsilon_{g_i})}{4n_i}}\bigg)^2 n_i}{2\sigma_i^2 + \frac{4}{3}\sqrt{\frac{\rho_v\psi[\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}]\log(T\epsilon_{g_i})}{4n_i} + \frac{\rho_v\psi\log(T\epsilon_{g_i})}{4n_i}}}\Bigg)$$
$$\leq \exp\Bigg(-\frac{\Big(\rho_v\psi[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1]\log(T\epsilon_{g_i})\Big)}{2\sigma_i^2 + \frac{2}{3}\sqrt{\rho_v\epsilon_{g_i}}}\Bigg)$$
$$\leq \exp\bigg(-\frac{3\rho_v\psi}{2}\bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\bigg)\log(T\epsilon_{g_i})\bigg)$$
$$\leq \exp\bigg(-\frac{3\rho_v T\epsilon_{g_i}}{16K\log K}\bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\bigg)\log(T\epsilon_{g_i})\bigg),$$

putting the value of $\psi = \frac{T\epsilon_{m_i}}{8K\log K}$

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

$$\mathbb{P}\bigg\{\hat{V}_i \geq \sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}\bigg\}$$
$$\leq \mathbb{P}\bigg\{\frac{1}{n_i}\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2 - (\hat{r}_i - r_i)^2 \geq \sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}\bigg\}$$
$$\leq \mathbb{P}\bigg\{\frac{\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}\bigg\}$$
$$\leq \mathbb{P}\bigg\{\frac{\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 +$$
$$\bigg(2\sqrt{\frac{\rho_v\psi[\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}]\log(T\epsilon_{g_i})}{4n_i} + \frac{\rho_v\psi\log(T\epsilon_{g_i})}{4n_i}}\bigg)\bigg\}$$
$$\leq \exp\bigg(-\frac{3\rho_v\psi}{2}\bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\bigg)\log(T\epsilon_{g_i})\bigg)$$
$$\leq \exp\bigg(-\frac{3\rho_v T\epsilon_{g_i}}{16K\log K\epsilon_{g_i}}\bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\bigg)\log(T\epsilon_{g_i})\bigg),$$

putting the value of $\psi = \frac{T\epsilon_{m_i}}{8K\log K}$

Similarly, the condition for the complementary event for the elimination case 4 holds such that $\mathbb{P}\{\hat{r}_i < r_i - 2s_i\} \leq 2\exp\Big(-\frac{3T\rho_v\epsilon_{g_i}}{16K\log K}\Big(\frac{\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}+1}{3\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}}\Big)\log(T\epsilon_{g_i})\Big)$.

Again summing the above expressions, 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 $4\exp\Big(-\frac{3T\rho_v\epsilon_{g_i}}{16K\log K}\Big(\frac{\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}+1}{3\sigma_i^2+\sqrt{\rho_v\epsilon_{g_i}}}\Big)\log(T\epsilon_{g_i})\Big)$.

Hence, for the i-th arm we can bound the probability of error till the event ξ_1 or ξ_2 by,

$$\mathbb{P}\{\xi_1\} + \mathbb{P}\{\xi_2\} \geq 1 - (\mathbb{P}\{|\hat{r}_i - r_i| > 2c_i\} + \mathbb{P}\{|\hat{r}_i - r_i| > 2s_i\})$$
$$\geq 1 - \bigg(\bigg(2\exp\bigg(-\frac{T\rho_\mu\epsilon_{m_i}}{2K\log K}\log(T\epsilon_{m_i})\bigg)\bigg)$$
$$+ 4\exp\bigg(-\frac{3T\rho_v\epsilon_{g_i}}{16K\log K}\bigg(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\bigg)\log(T\epsilon_{g_i})\bigg)\bigg)$$

Now, in the m_i-th round or in the g_i-th round we know that $\sqrt{\epsilon_{m_i}\rho_\mu} < \frac{\Delta_i}{2}$ or $\sqrt{\epsilon_{g_i}\rho_v} < \frac{\Delta_i}{2}$.

$$\mathbb{P}\{\xi_1\} + \mathbb{P}\{\xi_2\} \geq 1 - \Bigg(2\exp\Bigg(-\frac{T\rho_\mu\frac{\Delta_i^2}{4\rho_\mu}}{2K\log K}\log(T\frac{\Delta_i^2}{4\rho_\mu})\Bigg)$$
$$+ 4\exp\Bigg(-\frac{3T\rho_v\frac{\Delta_i^2}{4\rho_v}}{16K\log K}\bigg(\frac{\sigma_i^2 + \frac{\Delta_i}{2} + 1}{3\sigma_i^2 + \frac{\Delta_i}{2}}\bigg)\log(T\frac{\Delta_i^2}{4\rho_v})\Bigg)\Bigg)$$
$$\geq 1 - \bigg(2\exp\bigg(-\frac{T\Delta_i^2}{8K\log K}\log(2T\Delta_i^2)\bigg)$$
$$+ 4\exp\bigg(-\frac{3T\Delta_i^2}{16K\log K}\bigg(\frac{2\sigma_i^2 + \Delta_i + 2}{6\sigma_i^2 + \Delta_i}\bigg)\log(\frac{3}{4}T\Delta_i^2)\bigg)\bigg),$$

putting the values of ρ_μ and ρ_v.

Now, we know that $\mathbb{E}[\mathcal{L}(T)] \leq 1 - (\mathbb{P}\{\xi_1\} + \mathbb{P}\{\xi_2\})$. Also, we know from [Bubeck *et al.*, 2011] and [Auer and Ortner, 2010] that the function $x \in [0,1] \mapsto x\exp(-Cx^2)$ is decreasing on $[1/\sqrt{2C}, 1]$ for any $C > 0$. So, taking $C = \lfloor e/T \rfloor$ and putting $\min_{i\in A}\Delta_i = \Delta = \sqrt{\frac{K\log K}{T}} > \sqrt{\frac{e}{T}}, \forall i \in A$ and summing over all arms K and over all rounds $m = 0, 1, 2, .., \max\{m_i, g_i\}$ we get that,

$$\mathbb{E}[\mathcal{L}(T)] \leq \sum_{i=1}^{K}\sum_{m=0}^{\max\{m_i,g_i\}}\bigg\{\bigg(2\exp\bigg(-\frac{2T\Delta_i^2}{8K\log K}\log(2T\Delta_i^2)\bigg)$$
$$+ 4\exp\bigg(-\frac{3T\Delta_i^2}{64K\log K}\bigg(\frac{2\sigma_i^2 + \Delta_i + 2}{6\sigma_i^2 + \Delta_i}\bigg)\log(\frac{3}{4}T\Delta_i^2)\bigg)\bigg\}$$

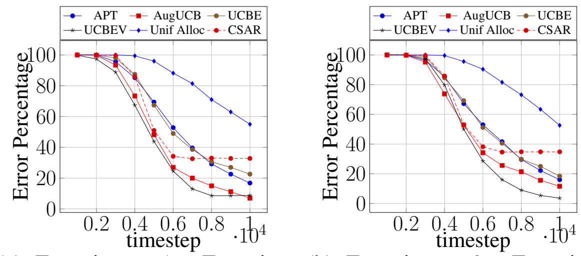
$$
\leq K \sum_{m=0}^{M} \Bigg\{ 2\exp\left(- \frac{T}{i \max_i \Delta_i^{-2}} . \frac{T \log(2K \log K)}{16K \log K} \right)
$$

$$
+ 4\exp\left(- \frac{6T}{i \max_i (6\sigma_i^2 + \Delta_i)\Delta_i^{-2}} . \frac{\log(\frac{3}{4}K \log K)}{64K \log K} \right) \Bigg\}
$$

$$
\leq K \left(\log_2 \frac{T}{e} + 1 \right) \Bigg\{ \exp\left(- \frac{T \log(2K \log K)}{16 H_2 K \log K} \right)
$$

$$
+ 2\exp\left(- \frac{T \log(\frac{3}{4}K \log K)}{64 H_2^{\sigma} K (\log K)} \right) \Bigg\}
$$

□



(a) Experiment 1: Experiment with Arithmetic Progression

(b) Experiment 2: Experiment with Geometric Progression

Figure 1: Experiments with thresholding bandit

5 Numerical Experiments

In this section we compare the empirical performance of AugUCB against APT, Uniform Allocation, CSAR, UCBE and UCBEV algorithm. The threshold τ is set at 0.5 for all experiments. Each algorithm is run independently 500 times for 10000 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 10000 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 $a = \frac{T-K}{H_1}$. Again, for UCBEV, following [Gabillon *et al.*, 2011], we modify it such that the exploration parameter $a = \frac{T-2K}{H_1^{\sigma}}$ where $H_1^{\sigma} = \sum_{i=1}^{K} \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}$. Then for each timestep $t = 1, 2, .., T$ we pull the arm that minimizes $\{|\hat{r}_i - \tau| - \sqrt{\frac{a}{n_i}}\}$, where n_i is the number of times the arm i is pulled till $t-1$ timestep and a is set as mentioned above for UCBE and UCBEV respectively. Also, APT is run with $\epsilon = 0.05$, which denotes the precision with which the algorithm suggests the best set of arms and we modify CSAR as per [Locatelli *et al.*, 2016] such that it behaves as a Successive Reject algorithm whereby it rejects the arm farthest from τ after each phase. For AugUCB we take $\psi = \frac{T\epsilon_m}{8K \log K}$, $\rho_\mu = \frac{1}{8}$ and $\rho_v = \frac{1}{3}$ as in Theorem 4.1.

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 $\sigma_{i=1:5}^2 = 0.5$ and $\sigma_{i=6:10}^2 = 0.6$. Then $\sigma_{i=11:100}^2$ is chosen uniform randomly between $0.38 - 0.42$. 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 whereas arms below τ have lower variance. The result is shown in Figure 1(a). In this experiment we see that UCBEV which has access to the problem complexity and is a variance-aware algorithm beats all other algorithm including UCBE which has access to the problem complexity but does not take into account the variance of the arms. AugUCB with the said parameters outperforms UCBE, APT and the other non variance-aware algorithms that we have considered.

The second experiment is conducted on a testbed of 100 arms with the means of first 10 arms set as Geometric Progression. The testbed involves Gaussian reward distribution with expected rewards of the arms as $r_{1:4} = 0.4 - (0.2)^{1:4}$, $r_5 = 0.45$, $r_6 = 0.55$ and $r_{7:10} = 0.6 + (0.2)^{5-(1:4)}$. The variances of the arms $11 - 100$ are set in the same way as in Experiment 1. AugUCB, APT, CSAR, 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, again we see that AugUCB beats APT, UCBE and all the non-variance aware algorithms with only UCBEV beating AugUCB.

6 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.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 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.

[Jamieson and Nowak, 2014] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–6. IEEE, 2014.

[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.