Thresholding Bandits with Augmented UCB

Author names withheld

Abstract

We propose the Augmented-UCB (AugUCB) algorithm for the thresholding bandit problem, which is an instance of the combinatorial fixed-budget pure-exploration stochastic multi-armed bandit setup. Our algorithm is based on arm elimination, employing both the mean and variance estimates. Theoretically, our algorithm provides a weaker guarantee (in terms of an upper bound on the expected loss) than APT [Locatelli *et al.*, 2016] and CSAR [Chen *et al.*, 2014]. However, through simulation experiments we establish that our algorithm, being variance-aware, performs better than APT and CSAR algorithms, particularly when a large number of arms are involved.

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*).

Pure exploration problems are unlike their traditional (exploration vs. exploitation) counterparts where the objective is to minimize the cumulative regret, which is the total loss incurred by the learner for not playing the optimal arm throughout the time horizon T. Instead, in the pure exploration setup the learning algorithm is provided with a threshold τ, and the objective, after exploring for T rounds, is to output all arms i whose r_i is above τ. Thus, the learning algorithm, until time T, can invest entirely on exploring the arms without being concerned about the loss incurred while exploring.

Formally, the problem we consider is the following. First, we define the set $S_\tau = \{i \in \mathcal{A} : r_i \geq \tau\}$. Note that, S_τ is the set of all arms whose reward mean is greater than τ. Let S_τ^c denote the complement of S_τ, i.e., $S_\tau^c = \{i \in \mathcal{A} : r_i < \tau\}$. Next, let $\hat{S}_\tau = \hat{S}_\tau(T) \subseteq \mathcal{A}$ denote the recommendation of the learning algorithm after T time units of exploration, while $\hat{S}_\tau^c$ denotes its complement. The performance of the learning agent is measured by the accuracy with which it can classify the arms into S_τ and S_τ^c after time horizon T. Equivalently, using $\mathbb{I}(E)$ to denote the indicator of an event E, the *loss* $\mathcal{L}(T)$ is defined as

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

Finally, the goal of the learning agent is to minimize the expected loss:

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

Note that the expected loss is simply the *probability of error*, that occurs either if a good arm is rejected or a bad arm is accepted as a good one.

The above TBP formulation has several applications, for instance, from areas ranging from anomaly detection and classification [Locatelli *et al.*, 2016] to industrial application. Particularly in industrial applications a learners objective is to choose (i.e., keep in operation) all machines whose productivity is above a threshold. Similarly, TBP finds applications in mobile communications [Audibert and Bubeck, 2010] where the users are to be allocated only those channels whose quality is above an acceptable threshold.

1.1 Related Work

Significant amount of literature is available on the stochastic MAB setting with respect to minimizing the cumulative regret. While the seminal work of [Robbins, 1952], [Thompson, 1933], and [Lai and Robbins, 1985] prove asymptotic lower bounds on the cumulative regret, the more recent work of [Auer *et al.*, 2002] propose the UCB1 algorithm that provides finite time-horizon guarantees. Subsequent work such as [Audibert and Bubeck, 2009] and [Auer and Ortner, 2010] have improved the upper bounds on the cumulative regret. The authors in [Auer and Ortner, 2010] have proposed a *round-based*[1] version of the UCB algorithm, referred to as UCB-Improved. Of special mention is the work of [Audibert *et al.*, 2009] where the authors have introduced a *variance-aware* UCB algorithm, referred to as UCB-V; it is shown that the algorithms that take into account variance estimation along with mean estimation tends to perform better than the algorithms that solely focuses on mean estimation, for instance, such as UCB1. For a more detail survey of literature on UCB algorithms, we refer the reader to [Bubeck and Cesa-Bianchi, 2012].

In this work we are particularly interested in *pure-exploration MABs*, where the focus in primarily on simple regret rather than the cumulative regret. 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. The pure exploration problem has been explored mainly under the following two settings:

1. Fixed Budget setting: Here the learning algorithm has to suggest the best arm(s) within a fixed time-horizon T, that is usually given as an input. The objective is to maximize the probability of returning the best arm(s). This is the scenario we consider in our paper. In [Audibert and Bubeck, 2010] the authors propose the UCBE and the Successive Reject (SR) algorithm, and prove simple-regret guarantees for the problem of identifying the single best arm. In the combinatorial fixed budget setup [Gabillon *et al.*, 2011] propose the Gap-E and Gap-EV algorithms that suggests, with high probability, the best m arms at the end of the time budget. Similarly, [Bubeck *et al.*, 2013] introduce the Successive Accept Reject (SAR) algorithm, which is an extension of the SR algorithm; SAR is a round based algorithm whereby at the end of each round an arm is either accepted or rejected (based on certain confidence conditions) until the top m arms are suggested at the end of the budget with high probability. A similar combinatorial setup was explored in [Chen *et al.*, 2014] where the authors propose the Combinatorial Successive Accept Reject (CSAR) algorithm, which is similar in concept to SAR but with a more general setup.

2. Fixed Confidence setting: In this setting the learning algorithm has to suggest the best arm(s) with a fixed confidence (given as input) with as fewer number of attempts as possible. The single best arm identification has been studied in [Even-Dar *et al.*, 2006], while for the combinatorial setup [Kalyanakrishnan *et al.*, 2012] have proposed the LUCB algorithm which, on termination, returns m arms which are at least ϵ close to the true top-m arms with probability at least $1-\delta$. For a detail 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 of [Locatelli *et al.*, 2016] Anytime Parameter-Free Thresholding (APT) algorithm comes up with an improved anytime guarantee than CSAR for the thresholding bandit problem.

[1]An algorithm is said to be *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.

1.2 Our Contribution

In this paper we propose the Algorithm AugUCB which is an 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 $H_1, H_2, H_{CSAR,2}, H_1^\sigma$ and H_2^σ signifies problem complexity and are defined in section 3. Theoretically, we can compare the first term (containing H_2) of our expected loss and see that for all $K \geq 4$, $H_2 \log(\frac{3}{16} K \log K) > (\log K) H_{CSAR,2} \geq H_1$ and hence our result is weaker than CSAR and APT. The term containing H_2^σ is comparable to the similar terms (containing H_1^σ) for the error probability of Gap-EV([Gabillon *et al.*, 2011] or UGapE-V([Gabillon *et al.*, 2012]) algorithm which we modify to perform in the TBP problem. The error probability of Gap-EV for single bandit multi-armed case is given by $6TK \exp(-\frac{1}{512}\frac{T-2K}{H_1^\sigma})$ where $\log(\frac{3}{16} K \log K) H_2^\sigma > H_1^\sigma$ and hence our algorithm is weaker with respect to Gap-EV for single multi-armed bandit scenario. But Gap-EV algorithm needs the complexity factor H_1^σ as input for optimal performance (which is not a realistic scenario) whereas AugUCB requires no such complexity factor as input.

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	$K^2 \exp(-\frac{T-K}{72\log(K)H_{CSAR,2}})$
AugUCB	$\exp\left(-\frac{T}{64H_2 a} + \log\left(K\left(\log_2 \frac{T}{e} + 1\right)\right)\right) +$ $\exp\left(-\frac{T}{4096 H_2^\sigma a} + \log\left(2K\left(\log_2 \frac{T}{e} + 1\right)\right)\right)$ where $a = \log(\frac{3}{16} K \log K)$

Empirically we show that for a large number of arms when the variance of the arms lying above τ are high, our algorithm performs better than all other algorithms, except the algorithm UCBEV (modified from Gap-EV for TBP) 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 3.1. 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 re-allocating 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 remainder of the paper is organized as follows. In section 2 we present our AugUCB algorithm. Section 3 contains our main theorem on expected loss, while section 4 contains the numerical experiments. We finally draw our conclusions in section 5.

2 Augmented UCB

Notations and assumptions: $\mathcal{A}$ denotes the set of arms, and $|\mathcal{A}| = K$ is the number of arms in $\mathcal{A}$. a generic arm is indexed by $i, j \in \mathcal{A}$. For arm $i \in \mathcal{A}$, we use r_i to denote the true mean of the distribution from which the rewards are sampled, while $\hat{r}_i(t)$ denotes the estimated mean at time t. Formally, using $n_i(t)$ to denote the number of times arm i has been pulled until time t, we have $\hat{r}_i(t) = \frac{1}{n_i(t)} \sum_{z=1}^{n_i(t)} X_{i,z}$, where $X_{i,z}$ is the reward sample received when arm i is pulled for the z-th time. For simplicity, whenever there is no confusion about the time index t, we simply neglect the denote $\hat{r}_i(t)$ and $\hat{r}_i$ and $n_i(t) = n_i$ whenever there is no confusion about the time index t (this nomenclature also holds for the following notation). Similarly, for arm i we use σ_i^2 to denote the true variance of the corresponding reward distribution, while $\hat{v}_i(t)$ is the estimated variance, i.e., $\hat{v}_i(t) = \frac{1}{n_i(t)} \sum_{z=1}^{n_i(t)} (X_{i,z} - \hat{r}_i)^2$. Let $\Delta_i = |\tau - r_i|$ and $\hat{\Delta}_i = |\tau - \hat{r}_i|$.

We assume 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 distributions supported on an interval of length less than 2. We will also assume that all rewards are bounded in $[0, 1]$.

Algorithm: In algorithm 1, hence referred to as AugUCB, we have two exploration parameters, ρ_μ and ρ_v which are the arm elimination parameters. ψ_m 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 $\mathcal{A}$. We divide the entire budget T into rounds/phases as like UCB-Improved, CCB, SAR and CSAR. At every timestep AugUCB checks for arm elimination conditions and update parameters after finishing a round. As suggested by [Liu and Tsuruoka, 2016] to make AugUCB an 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, $\min_{i \in B_m} \left\{ |\hat{r}_i - \tau| - 2\sqrt{\frac{\rho_v \psi_m \hat{V}_i \log(T\epsilon_m)}{4n_i} + \frac{\rho_v \psi_m \log(T\epsilon_m)}{4n_i}} \right\}$ 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 and ρ_v we can fine tune the exploration. The choice of exploration factor $\psi_m = \frac{T\epsilon_m}{(\log(\frac{3}{16}K\log K))^2}$ 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 in T which is suited for minimizing cumulative regret.

Algorithm 1 AugmentedUCB

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

Initialization: Set $B_0 := \mathcal{A}$, $M = \left\lfloor \frac{1}{2}\log_2 \frac{T}{e} \right\rfloor$, $m := 0$, $\epsilon_0 := 1$, $\psi_0 = \frac{T\epsilon_0}{(\log(\frac{3}{16}K\log K)^2}$, $\ell_0 = \left\lceil \frac{2\psi\log(T\epsilon_0)}{\epsilon_0} \right\rceil$ 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

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

$$\text{where } c_i = \sqrt{\frac{\rho_\mu \psi_m \log(T\epsilon_m)}{2n_i}}$$

Arm Elimination by Mean and Variance Estimation

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

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

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

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

Reset Parameters

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

$B_{m+1} := B_m$

$\psi_{m+1} = \frac{T\epsilon_{m+1}}{(\log(\frac{3}{16}K\log K))^2}$

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

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

$m := m + 1$

end if

end for

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

---

3 Main Results

3.1 Problem Complexity

We define problem complexity as,

$$H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}, \; H_2 = \min_{i \in \mathcal{A}} \frac{i}{\Delta_{(i)}^2}, \text{ where } \Delta_i = |r_i - \tau| \text{ and}$$

$\Delta_{(i)}$ is an increasing ordering of Δ_i

H_1 and H_2 is same as the problem complexity defined in [Locatelli *et al.*, 2016] for the thresholding bandit problem while $H_{CSAR,2} = \max_i \frac{i}{\Delta^2_{(i)}}$ is defined in [Chen *et al.*, 2014]. Also we know from [Locatelli *et al.*, 2016] that,

$$H_2 \leq H_1 \leq \log(2K)H_2 \text{ and } H_1 \leq \log(K)H_{CSAR,2}$$

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

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

$$H_2^\sigma = \min_{i\in\mathcal{A}} i\tilde{\Delta}_{(i)}^{-2}, \text{ where } \tilde{\Delta}_i^{-2} = \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}$$

which also gives us that $H_2^\sigma \leq H_1^\sigma \leq \log(2K)H_2^\sigma$.

3.2 Theorem 1

Theorem 3.1. *For $K \geq 4$, with $\rho_\mu = \frac{1}{8}$ and $\rho_v = \frac{1}{3}$, the expected loss of the AugUCB algorithm is given by,*

$$\mathbb{E}[\mathcal{L}(T)] \leq K\left(\log_2 \frac{T}{e} + 1\right)\left\{\exp\left(-\frac{T}{64H_2 a}\right) + 2\exp\left(-\frac{T}{4096H_2^\sigma a}\right)\right\}$$

where $a = \log(\frac{3}{16}K\log K)$

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 $\ell_m = \left\lceil \frac{2\psi_m \log(T\epsilon_m)}{\epsilon_m}\right\rceil$, B_m is the set of all surviving arms and let $a = (\log(\frac{3}{16}K\log K))$.

Let $c_i = \sqrt{\frac{\rho_\mu \psi_m \log(T\epsilon_m)}{2n_i}}$ denote the confidence interval, where n_i is the number of times an arm i is pulled. Let $\mathcal{A}' = \{i \in \mathcal{A}|\Delta_i \geq b\}$, for $b \geq \sqrt{\frac{e}{T}}$. Define $m_i = \min\{m|\sqrt{\rho_\mu\epsilon_m} < \frac{\Delta_i}{2}\}$.

Let $s_i = \sqrt{\frac{\rho_v\psi_g\hat{V}_i\log(T\epsilon_g)}{4n_i} + \frac{\rho_v\psi_g\log(T\epsilon_g)}{4n_i}}$ 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,

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

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

So, ξ_1 and ξ_2 signifies the event 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,

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

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

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

Again, for $i \in \mathcal{A}'$ 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 $i \in \mathcal{A}'$ 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),

$$\mathbb{P}\{\hat{r}_i > r_i + 2c_i\} \leq \exp(-4c_i^2 n_i)$$
$$\leq \exp(-8 * \frac{\rho_\mu\psi_{m_i}\log(T\epsilon_{m_i})}{2n_i} * n_i)$$
$$\leq \exp\left(-4\rho_\mu\psi_{m_i}\log(T\epsilon_{m_i})\right)$$
$$\leq \exp\left(-\rho_\mu\frac{T\epsilon_{m_i}}{32a^2}\log(T\epsilon_{m_i})\right),$$

putting the value of $\psi_{m_i} = \frac{T\epsilon_{m_i}}{128(\log(\frac{3}{16}K\log K))^2}$

Similarly for the condition in (2), $\mathbb{P}\{\hat{r}_i < r_i - 2c_i\} \leq \exp\left(-\frac{T\rho_\mu\epsilon_{m_i}}{32a^2}\log(T\epsilon_{m_i})\right)$.

Summing the above two expressions, the probability that arm i is not eliminated on or before m_i-th is $\left(2\exp\left(-\frac{T\rho_\mu\epsilon_{m_i}}{32a^2}\log(T\epsilon_{m_i})\right)\right)$.

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

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

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

$$s_i = \sqrt{\frac{\rho_v\psi_{g_i}\hat{V}_i\epsilon_{g_i}\log(T\epsilon_{g_i})}{4n_i} + \frac{\rho_v\psi_{g_i}\log(T\epsilon_{g_i})}{4n_i}}$$

$$\leq \sqrt{\frac{\rho_v \psi_{g_i} \epsilon_{g_i} \log(T\epsilon_{g_i})}{4*2\log(\psi_{g_i} T\epsilon_{g_i})} + \frac{\rho_v \psi_{g_i} \epsilon_{g_i} \log(T\epsilon_{g_i})}{4*2\psi_{g_i} \log(T\epsilon_{g_i})}}, \text{ as } \hat{V}_i \in [0,1].$$

$$\leq \sqrt{\frac{\rho_v \epsilon_{g_i}}{8} + \frac{\rho_v \epsilon_{g_i}}{8}} \leq \frac{\sqrt{\rho_v \epsilon_{g_i}}}{2} < \frac{\Delta_i}{4}, \text{ as } \rho_v \in (0,1].$$

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

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

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

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

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

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

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

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

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

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

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

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

$$\leq \frac{\sqrt{\rho_v \epsilon_{g_i} [\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1]}}{2} \leq \sqrt{\rho_v \epsilon_{g_i}}$$

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

$$\mathbb{P}\left\{\hat{r}_i > r_i + \left(2\sqrt{\frac{\rho_v \psi_{g_i} [\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1] \log(T\epsilon_{g_i})}{4n_i}}\right)\right\}$$

$$\leq \exp\left(-\frac{\left(2\sqrt{\frac{\rho_v \psi_{g_i} [\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1] \log(T\epsilon_{g_i})}{4n_i}}\right)^2 n_i}{2\sigma_i^2 + \frac{4}{3}\sqrt{\frac{\rho_v \psi_{g_i} [\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1] \log(T\epsilon_{g_i})}{4n_i}}}\right)$$

$$\leq \exp\left(-\frac{\left(\rho_v \psi_{g_i} [\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1] \log(T\epsilon_{g_i})\right)}{2\sigma_i^2 + \frac{2}{3}\sqrt{\rho_v \epsilon_{g_i}}}\right)$$

$$\leq \exp\left(-\frac{3\rho_v \psi_{g_i}}{2}\left(\frac{\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}}\right) \log(T\epsilon_{g_i})\right)$$

$$\leq \exp\left(-\frac{3\rho_v T\epsilon_{g_i}}{256a^2}\left(\frac{\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}}\right) \log(T\epsilon_{g_i})\right),$$

putting the value of $\psi_{g_i} = \dfrac{T\epsilon_{g_i}}{128(\log(\frac{3}{16}K\log K))^2}$

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

$$\mathbb{P}\left\{\hat{V}_i \geq \sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}\right\}$$

$$\leq \mathbb{P}\left\{\frac{1}{n_i}\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2 - (\hat{r}_i - r_i)^2 \geq \sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}\right\}$$

$$\leq \mathbb{P}\left\{\frac{\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}\right\}$$

$$\leq \mathbb{P}\left\{\frac{\sum_{t=1}^{n_i}(x_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \left(2\sqrt{\frac{\rho_v \psi_{g_i} [\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}] \log(T\epsilon_{g_i})}{4n_i} + \frac{\rho_v \psi_{g_i} \log(T\epsilon_{g_i})}{4n_i}}\right)\right\}$$

$$\leq \exp\left(-\frac{3\rho_v \psi_{g_i}}{2}\left(\frac{\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}}\right) \log(T\epsilon_{g_i})\right)$$

$$\leq \exp\left(-\frac{3\rho_v T\epsilon_{g_i}}{256a^2}\left(\frac{\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v \epsilon_{g_i}}}\right) \log(T\epsilon_{g_i})\right),$$

putting the value of $\psi_{g_i} = \dfrac{T\epsilon_{g_i}}{128(\log(\frac{3}{16}K\log K))^2}$

Similarly, the condition for the complementary event for the elimination case 4 holds such that $\mathbb{P}\{\hat{r}_i < r_i - 2s_i\} \leq 2\exp\left(-\frac{3T\rho_v\epsilon_{g_i}}{256a^2}\left(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\right)\log(T\epsilon_{g_i})\right)$.

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 $4\exp\left(-\frac{3T\rho_v\epsilon_{g_i}}{256a^2}\left(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\right)\log(T\epsilon_{g_i})\right)$.

Hence, for the i-th arm we can bound the probability of it getting eliminated till $\max\{m_i, g_i\}$-th round by,

$\mathbb{P}\{i \in \mathcal{A}'$ getting eliminated on or before round $\max\{m_i, g_i\}\}$

$$\geq 1 - (\mathbb{P}\{|\hat{r}_i - r_i| > 2c_i\} + \mathbb{P}\{|\hat{r}_i - r_i| > 2s_i\})$$

$$\geq 1 - \left(\left(2\exp\left(-\frac{T\rho_\mu\epsilon_{m_i}}{32a^2}\log(T\epsilon_{m_i})\right)\right) + 4\exp\left(-\frac{3T\rho_v\epsilon_{g_i}}{256a^2}\left(\frac{\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}} + 1}{3\sigma_i^2 + \sqrt{\rho_v\epsilon_{g_i}}}\right)\log(T\epsilon_{g_i})\right)\right)$$

Now, in the m_i-th round or in the g_i-th round we know that $\frac{\Delta_i}{4} \leq \sqrt{\epsilon_{m_i}\rho_\mu} < \frac{\Delta_i}{2}$ or $\frac{\Delta_i}{4} \leq \sqrt{\epsilon_{g_i}\rho_v} < \frac{\Delta_i}{2}$.

$\mathbb{P}\{i \in \mathcal{A}'$ getting eliminated on or before round $\max\{m_i, g_i\}\}$

$$\geq 1 - \left(2\exp\left(-\frac{T\rho_\mu\frac{\Delta_i^2}{16\rho_\mu}}{32a^2}\log(T\frac{\Delta_i^2}{16\rho_\mu})\right) + 4\exp\left(-\frac{3T\rho_v\frac{\Delta_i^2}{16\rho_v}}{256a^2}\left(\frac{\sigma_i^2 + \frac{\Delta_i}{4} + 1}{3\sigma_i^2 + \frac{\Delta_i}{4}}\right)\log(T\frac{\Delta_i^2}{16\rho_v})\right)\right)$$

$$\geq 1 - \left(2\exp\left(-\frac{T\Delta_i^2}{64a}\log(\frac{T\Delta_i^2}{2})\right)\right.$$
$$\left. + 4\exp\left(-\frac{3T\Delta_i^2}{4096a^2}\left(\frac{4\sigma_i^2+\Delta_i+4}{12\sigma_i^2+\Delta_i}\right)\log(\frac{3}{16}T\Delta_i^2)\right)\right),$$

putting the values of ρ_μ and ρ_v.

Again, $\mathbb{P}\{\xi_1 \cup \xi_2\} \geq 1 - \sum_{i=1}^{K}\sum_{m=0}^{\max\{m_i,g_i\}}\mathbb{P}\{i \in \mathcal{A}'$ getting eliminated on or before round $\max\{m_i, g_i\}\}$. Also, $\mathbb{E}[\mathcal{L}(T)] \leq 1 - \mathbb{P}\{\xi_1 \cup \xi_2\}$. 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 $C = \lfloor\sqrt{e/T}\rfloor$ and putting $\min_{i\in\mathcal{A}}\Delta_i = \Delta = \sqrt{\frac{K\log K}{T}} > \sqrt{\frac{e}{T}}, \forall i \in \mathcal{A}$ we get that,

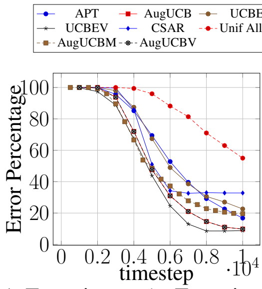
$$\mathbb{E}[\mathcal{L}(T)] \leq \sum_{i=1}^{K}\sum_{m=0}^{\max\{m_i,g_i\}}\left\{\left(2\exp\left(-\frac{T\Delta_i^2\log(\frac{T\Delta_i^2}{2})}{64a^2}\right)\right.\right.$$
$$\left.\left. + 4\exp\left(-\frac{3T\Delta_i^2}{4096a^2}\left(\frac{4\sigma_i^2+\Delta_i+4}{12\sigma_i^2+\Delta_i}\right)\log(\frac{3}{16}T\Delta_i^2)\right)\right\}\right.$$
$$\leq K\sum_{m=0}^{M}\left\{2\exp\left(-\frac{T}{\min_i i\Delta_{(i)}^{-2}}.\frac{\log(\frac{1}{2}K\log K)}{64a^2}\right)\right.$$
$$\left. + 4\exp\left(-\frac{12T\Delta_i^2}{(12\sigma_i+12\Delta_i)}.\frac{\log(\frac{3}{16}K\log K)}{4096a^2}\right)\right\}$$
$$\leq K\left(\log_2\frac{T}{e}+1\right)\left\{\exp\left(-\frac{T\log(\frac{1}{2}K\log K)}{64H_2a^2}\right)\right.$$
$$\left. + 2\exp\left(-\frac{T\Delta_i^2\log(\frac{3}{16}K\log K)}{4096(\sigma_i+\sqrt{\sigma_i^2+(16/3)\Delta_i})a^2}\right)\right\}$$
$$\leq K\left(\log_2\frac{T}{e}+1\right)\left\{\exp\left(-\frac{T\log(\frac{1}{2}K\log K)}{64H_2(\log(\frac{3}{16}K\log K))^2}\right)\right.$$
$$\left. + 2\exp\left(-\frac{T\log(\frac{3}{16}K\log K)}{4096\min_i i\tilde{\Delta}_{(i)}^{-2}(\log(\frac{3}{16}K\log K))^2}\right)\right\}$$
$$\leq K\left(\log_2\frac{T}{e}+1\right)\left\{\exp\left(-\frac{T}{64H_2(\log(\frac{3}{16}K\log K))}\right)\right.$$
$$\left. + 2\exp\left(-\frac{T}{4096H_2^\sigma(\log(\frac{3}{16}K\log K))}\right)\right\}$$

□

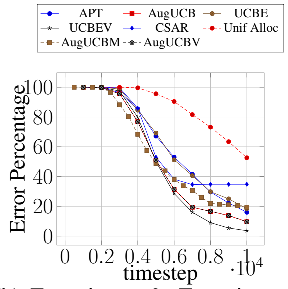
4 Numerical Experiments

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 $a = \frac{T-2K}{H_1^\sigma}$. 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 $\rho_\mu = \frac{1}{8}$ and $\rho_v = \frac{1}{3}$ as in Theorem 3.1. Also we run AugUCBM with arm elimination just by mean estimation and AugUCBV with arm elimination just by variance estimation. For AugUCBM, at every timestep we pull arm that minimizes $i \in \arg\min_{j\in B_m}\left\{|\hat{r}_j - \tau| - 2c_j\right\}$ while for AugUCBV we pull arm that minimizes $i \in \arg\min_{j\in B_m}\left\{|\hat{r}_j - \tau| - 2s_j\right\}$. 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 $\sigma_{1:5}^2 = 0.5$ and $\sigma_{6:10}^2 = 0.6$. Then $\sigma_{11:100}^2$ 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. AugUCBM with just mean estimation performs worse than AugUCB or AugUCBV, which have a matching performance in this setup.

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 all the arms and $r_{11:100}$ 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.



(a) Experiment 1: Experiment with Arithmetic Progression



(b) Experiment 2: Experiment with Geometric Progression

Figure 1: Experiments with thresholding bandit

5 Conclusion and Future work

From a theoretical viewpoint, we conclude that considering only arm elimination by mean estimation, the expected loss AugUCB is more than APT and CSAR while considering arm elimination by both mean and variance estimation we see that the expected loss of AugUCB is more than only UCBEV (which has access to problem complexity). From the numerical experiments on settings with large number of arms with different mean and variance, we observed that AugUCB outperforms all the non-variance aware algorithms. It would be interesting future research to come up with an anytime version of AugUCB algorithm. This is also the first paper to apply elimination by variance estimation in the TBP problem by modifying UCB-Improved and CCB algrithms.

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 non-stochastic 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.

6 Appendix(We will comment this out later)

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

$$H_2^\sigma = \min_{i \in \mathcal{A}} i\tilde{\Delta}_{(i)}^{-2}, \text{ where } \tilde{\Delta}_i^{-2} = \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}$$

We know that $\sigma_i \in [0,1], \forall i \in \mathcal{A}$ and $\Delta_i \in [0,1], \forall i \in \mathcal{A}$ and so $\sigma_i^2 \leq \sigma_i$ and $\sqrt{\Delta_i} \geq \Delta_i$.

$$\begin{aligned}
(3\Delta_i^2).\left(\frac{4\sigma_i^2 + \Delta_i + 4}{12\sigma_i^2 + \Delta_i}\right) &> \left(\frac{12\Delta_i^2}{12\sigma_i^2 + \Delta_i}\right) \\
&> \left(\frac{12\Delta_i^2}{12\sigma_i^2 + 12\Delta_i}\right) \\
&> \left(\frac{\Delta_i^2}{\sigma_i + \Delta_i}\right) \\
&> \left(\frac{\Delta_i^2}{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}\right) \\
&> \left(\frac{1}{\min_i i\tilde{\Delta}_i^2}\right)
\end{aligned}$$

Now, from [Audibert and Bubeck, 2010] we know that,

$$\begin{aligned}
\sum_{i=1}^{K} \tilde{\Delta}_i^{-2} = \tilde{\Delta}_{(2)}^{-2} + \sum_{i=2}^{K} \frac{1}{i} i\tilde{\Delta}_{(i)}^{-2} &\leq \bar{\log} K \min_i i\tilde{\Delta}_{(i)}^{-2} \\
&\leq \log(2K) H_2^\sigma, \text{ as } \bar{\log} K \leq \log(2K)
\end{aligned}$$

So, $H_2^\sigma \leq H_1^\sigma \leq \log(2K) H_2^\sigma$

Regarding union bound

$$\begin{aligned}
\mathbb{P}\{\xi_1 \cup \xi_2\} &= \mathbb{P}\{\xi_1\} + \mathbb{P}\{\xi_2\} - \mathbb{P}\{\xi_1 \cap \xi_2\} \\
&\leq \mathbb{P}\{\xi_1\} + \mathbb{P}\{\xi_2\}
\end{aligned}$$

So,

$$1 - \mathbb{P}\{\xi_1 \cup \xi_2\} \geq 1 - \mathbb{P}\{\xi_1\} + \mathbb{P}\{\xi_2\} \geq \mathbb{E}[\mathcal{L}(T)]$$