Solving Multi-Armed Bandits by Upper Confidence Bound Algorithms

Fan-Keng Sun, Chen-Hao Hsiao, Chi-Hsin Lo

b03901056, b03901030, b03901070

Abstract

The most simple example of the exploration versus exploitation dilemma is the stochastic multi-armed bandits (MAB) problem. Several algorithms have been proposed to tackle MAB problem, among them, algorithms based on upper confidence bound are the most successful and widely-used. Since the first paper about UCB [1] came out in 2002, there have been an ongoing research [2, 3, 4, 5, 6] in improving the strategy to obtain a lower regret bound. In this paper, we survey starting from the original UCB [1], to improved versions [2, 3], and end at the state-of-the-art method [4]. We also introduce the lower bound for a family of algorithms (*consistent* algorithms defined in [7]) and show the optimality of KL-UCB [8] in special case.

1 Introduction

Multi-armed bandits (MAB) problem is a form of online learning, but the learner only receives partial information about the environment. Specifically, in a K-armed bandits problem, there are K gambling machines (i.e. the arm of bandit), and in every trial t, the learner will choose an action (pull an arm) i_t, then the learner can observe the reward associated with only the pulled arm. The objective of the learner is to maximize the reward over the sequence of trials, or in other words, minimize the gap between the learner's decisions and the best decisions.

Although the original motivation of [9] for studying bandit problems came from clinical trials, where doctors will have to choose one medicine for a client and then observe the effects, there are much more applications of bandit problem in modern world. For example, companies have to decide what advertisement to display on the website for a sequential flow of customers to maximize overall click-through rate. Computer game-playing is another field where algorithms for bandits are successfully implemented.

There are three main formulations of MAB problem: stochastic, adversarial (non-stochastic) and Markovian. In the stochastic MAB setting, each arm is assumed to be represented by a fixed (but unknown) probability distribution, and reward are drawn i.i.d. from the distribution. However, in adversarial MABs, no probabilistic assumptions can be made on any arms. Instead, each arm tries to play against the learner. At last, the rewards from every arm follow a Markov process in the Markovian MABs, which is closely related to reinforcement learning. For these three different settings, three different playing strategies have been shown to effectively solve the problems: the UCB and its variants in the stochastic case, the Exp3 randomized algorithm [10] in the adversarial case, and the Gittins indices [11] in the Markovian case. Nevertheless, in this survey, we will be focusing on stochastic MABs only.

In stochastic MAB, the learner will faced the so-called exploration v.s. exploitation dilemma: shall it continue to select the best arm observed so far (exploitation) or rather probe other arms further (exploration). If the learner decides to exploit, then it takes the risk that its observation is inaccurate (i.e. optimal arm is underestimated). On the other hand, exploration may just be a waste of time. How to balance between exploitation and exploration is the main issue in solving MAB problem.

2 Problem Formulation

Stochastic MAB problem is parameterized by the following:

- $A = \{1, \ldots, K\}$: the set of arms, and $K = |A|$ is the number of arms.
- $\boldsymbol{X} = \{X_1, \ldots, X_K\}$: X_i is i.i.d random variable representing the reward from arm i, and distinct arms are independent. $X_i \in [0, 1]$ if not specified.
- $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$: $\mu_i = \mathbb{E}[X_i]$ is the mean reward of arm i, and $\mu^* = \max\{\boldsymbol{\mu}\}$ is the best mean reward, $i^* \in \{i | \mu_i = \mu^*\}$ is any optimal arm. Also, σ_i is the standard deviation of X_i.
- T: the number of trials (or budgets).

The learner will go through the following process

Algorithm 1: Solving MAB problems for K-amrs in T trials

for $t \coloneqq 1$ **to** T **do**
  Learner pulls an arm $i_t \in [K]$.
  Environment selects a reward vector $\boldsymbol{x_t} = \{x_{1,t}, \ldots, x_{K,t}\}$.
  Learner observes $x_{i_t,t}$.

while trying to minimize the (cumulative) regret R_T after T trials, which is defined as:

$$R_T = \mu^* T - \sum_{i=1}^{K} \mu_i N_{i,T},$$

where $N_{i,T}$ is the number of times the learner has chosen arm i up to trial T. The expected (cumulative) regret after trial T can be written as:

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_{i,T}], \tag{1}$$

where $\Delta_i = \mu^* - \mu_i$ is the gap between the means of optimal arm and the chosen arm.

3 Upper Confidence Bound Algorithms

How to balance between exploration and exploitation? Something better that ϵ-greedy [1]? An easy first thought is to use empirical mean reward $\hat{\mu}_i^n$ as an estimation of true mean reward μ_i for each arm i after pulling each arm at least once, where $\hat{\mu}_i^n$ is defined as the observed mean reward after pulling arm i n times. When the context is clear, we will drop the n in $\hat{\mu}_i^n$. Imagine if we pull each arm sufficiently large times, then we have sufficiently high confidence on the empirical mean. On the other hand, if an arm is pulled only a few times, then we have high uncertainty about its mean reward. This observation is the intuition behind upper confidence bound (UCB), which is a strategy based on "optimism in the face of uncertainty". We want a formal way to give a high confidence bound on true mean reward μ_i given the value of empirical mean reward $\hat{\mu}_i$ and the uncertainty of an arm i, which can be considered inverse proportion to the number of pulls $N_{i,t-1}$ on arm i. This can be accomplished by use of appropriate concentration inequalities.

Generally speaking, we can now bound the probability in trial t that (1) a sub-optimal arm i is over-estimated as $\hat{\mu}_i - \mu_i \geq \epsilon$, and (2) a optimal arm i^* is under-estimated as $\mu_{i^*} - \hat{\mu}_{i^*} \geq \epsilon$, where ϵ is a fixed value. By integrating the bound ϵ into algorithms, UCB-based algorithms are easy to understand and implement while successfully pave a road to solve the exploration v.s. exploitation dilemma.

Algorithm	Year	Regret Upper Bound
UCB1 [1]	2002	$O\left(\frac{K \log T}{\Delta}\right)$
improved-UCB [3]	2010	$O\left(\frac{K \log(T\Delta^2)}{\Delta}\right)$
UCBV [2]	2009	$O\left(\frac{K\sigma^2_{\max} \log T}{\Delta}\right)$
EUCBV [4]	2018	$O\left(\frac{K\sigma^2_{\max} \log(\frac{T\Delta^2}{K})}{\Delta}\right)$

Table 1: Comparison of UCB-based algorithms.

The first UCB-based algorithm is UCB1, which is proposed in [1], with a regret bound of $O(\frac{K \log T}{\Delta})$ will be introduced in section 4, where $\Delta = \min_{i:\Delta_i>0} \Delta_i$. In this paper, the base of log is e if not specified. Then, improved-UCB [3] uses a round-based procedure with an elimination phase to achieve a better bound $(\frac{K \log(T\Delta^2)}{\Delta})$, and will be discussed in section 5. Next, in section 6, UCB-variance (UCBV) considers variance and has a regret bound of $(\frac{K\sigma^2_{\max} \log T}{\Delta})$, where $\sigma_{\max} = \max_{i \in A} \sigma_i$. Finally, the latest efficient-UCBV (EUCBV), whose regret bound is $O(\frac{K\sigma^2 \log \frac{T\Delta^2}{K}}{\Delta})$, is described in section 7. Notice that $\Delta_i \leq 1$, and $\sigma_i \leq \frac{1}{4}$, since by convention $X_i \in [0, 1]$. The comparison of these algorithm is listed in table 1.

4 UCB1

Algorithm 2: UCB1 algorithm

Input: A stochastic MAB problem as defined in section 2.

for $t := 1$ **to** T **do**

Arm selection:

If $t \leq K$, select arm $i_t = t$. (Initialization).

If $t > K$, select arm $i_t \in \arg\max_{i \in A}(\hat{\mu}_i + \sqrt{\frac{2 \log T}{N_{i,t-1}}})$.

end

Theorem 1. *The upper bound total expected regret of the UCB1 algorithm until trail T is*

$$8 \sum_{i \in A, \Delta_i > 0} \left(\frac{\log T}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{3}\right) \sum_{i=1}^{K} \Delta_i \tag{2}$$

Before we prove the theorem, we would like to introduce the following lemma that captures the intuition of pulling a sub-optimal arm.

Lemma 2. *Fix any $i : \Delta_i > 0$. If UCB1 algorithm selects the arm i in trial t ($i_t = i$), then at least one of the following holds:*

(a) $\hat{\mu}_{i^*} \leq \mu^* - \sqrt{\frac{2 \log t}{N_{i^*,t-1}}}$

(b) $\hat{\mu}_i \geq \mu_i + \sqrt{\frac{2 \log t}{N_{i,t-1}}}$

(c) $N_{i,t-1} \leq \frac{8 \log T}{\Delta_i^2}$

Proof of Lemma 2: Prove by contradiction. If (a), (b), and (c) are all false, we have:

$$\begin{aligned}
\hat{\mu}_{i^*} + \sqrt{\frac{2\log t}{N_{i^*,t-1}}} &> \mu^* && \text{since (a) is false} \\
&= \mu_i + \Delta_i \\
&> \mu_i + \sqrt{\frac{8\log T}{N_{i,t-1}}} && \text{since (c) is false} \\
&\geq \mu_i + \sqrt{\frac{8\log t}{N_{i,t-1}}} \\
&> \hat{\mu}_i + \sqrt{\frac{2\log t}{N_{i,t-1}}} && \text{since (b) is false}
\end{aligned}$$

Which contradicts the fact that $i_t = i$. □

The lemma indicates that if we pull arm i in trial t, either we have not pulled the arm sufficiently many times, or we have been unlucky with the random draws of rewards and have a bad confidence interval constructed for either the pulled arm or an optimal arm.

Proof of Theorem 1: Since $\mathbb{E}[R_T] = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_{i,T}]$, we only need to bound $\mathbb{E}[N_{i,T}]$. Set $t_0 = \lceil \frac{8\log T}{\Delta_i^2} \rceil$ and let $\mathbb{1}(\cdot)$ denotes the indicator function, we then have:

$$\begin{aligned}
\mathbb{E}[N_{i,T}] &= \mathbb{E}[1 + \sum_{t=K+1}^{T} \mathbb{1}\{i_t = i\}] \\
&\leq \mathbb{E}[t_0 + \sum_{t=K+1}^{T} \mathbb{1}\{i_t = i, N_{i,t-1} \geq t_0\}] \\
&= t_0 + \sum_{t=K+1}^{T} \mathbb{P}(i_t = i, N_{i,t-1} \geq t_0) \\
&\leq t_0 + \sum_{t=K+1}^{T} \left(\mathbb{P}(\hat{\mu}_{i^*} \leq \mu^* - \sqrt{\frac{2\log t}{N_{i^*,t-1}}}) + \mathbb{P}(\hat{\mu}_i \geq \mu_i - \sqrt{\frac{2\log t}{N_{i,t-1}}}) \right) && \text{By Lemma 2} \\
&\leq t_0 + \sum_{t=K+1}^{T} \left(\mathbb{P}(\bigcup_{s=1}^{t-1} \{\hat{\mu}_{i^*} \leq \mu^* - \sqrt{\frac{2\log t}{s}}\}) + \mathbb{P}(\bigcup_{s'=1}^{t-1} \{\hat{\mu}_i \geq \mu_i - \sqrt{\frac{2\log t}{s'}}\}) \right) \\
&\leq t_0 + \sum_{t=1}^{T} \sum_{s=1}^{t} \sum_{s'=1}^{t} \left(\mathbb{P}(\hat{\mu}_{i^*} \leq \mu^* - \sqrt{\frac{2\log t}{s}}) + \mathbb{P}(\hat{\mu}_i \geq \mu_i - \sqrt{\frac{2\log t}{s'}}) \right)
\end{aligned}$$

By Chernoff-Hoeffding bound[12] we have:

$$\mathbb{P}(\hat{\mu}_{i^*} \leq \mu^* - \sqrt{\frac{2\log t}{s}}) \leq e^{-2s\frac{2\log t}{s}} = t^{-4}$$

and

$$\mathbb{P}(\hat{\mu}_i \geq \mu_i - \sqrt{\frac{2\log t}{s'}}) \leq e^{-2s'\frac{2\log t}{s'}} = t^{-4}$$

So

$$\begin{aligned}\mathbb{E}[N_{i,T}] &\leq t_0 + \sum_{t=1}^{T}\sum_{s=1}^{t}\sum_{s'=1}^{t} 2t^{-4} \\ &\leq t_0 + 2\sum_{t=1}^{T} t^{-2} \\ &\leq \frac{8\log T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}\end{aligned}$$

Summing over i we get:

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_{i,T}] \leq 8 \sum_{i \in A, \Delta_i > 0} \left(\frac{\log T}{\Delta_i}\right) + \left(1 + \frac{\pi^2}{3}\right) \sum_{i=1}^{K} \Delta_i \quad \square$$

5 Improved-UCB

In [3], based on UCB1, the authors designed an improved strategy as shown in Algorithm 3. Imagine if the learner had direct access to all Δ_i, then on can modify the confidence interval of UCB1 from $\sqrt{\frac{2\log T}{N_{i,t-1}}}$ to $\sqrt{\frac{2\log(T\Delta_i^2)}{N_{i,t-1}}}$, and the proof of the claimed bound of improved-UCB would be straightforward.

However, no Δ_i is known to the learner. Thus, in improved-UCB, the algorithm tries to guess the values of Δ_i by $\tilde{\Delta}$, which is initialized to 1 and halved each time the confidence intervals ($\sqrt{\frac{2\log(T\Delta_i^2)}{N_{i,t-1}}}$) become shorter thant $\tilde{\Delta}$. In addition, the algorithm also eliminated arms that perform bad enough. Similar algorithm were already proposed in [13], but it focuses on PAC analysis, instead of regret bound in [3].

Algorithm 3: Improved-UCB algorithm

Input: A stochastic MAB problem as defined in section 2.
$\tilde{\Delta}_0 \coloneqq 1$, and $B_0 \coloneqq A$
for round $m \coloneqq 0$ **do**
Arm selection:
If $|B_m| > 1$, choose each arm in B_m until the total number of times it has been chosen is

$$n_m \coloneqq \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil. \tag{3}$$

Otherwise choose the single arm in B_m until step T is reached.
Arm elimination:
Delete all arms i from B_m for which

$$\left(\hat{\mu}_i + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}\right) < \max_{j \in B_m} \left\{\hat{\mu}_j - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}\right\} \tag{4}$$

to obtain B_{m+1}.
Set $\tilde{\Delta}_{m+1}$:
$\tilde{\Delta}_{m+1} \coloneqq \frac{\tilde{\Delta}_m}{2}$
end

Theorem 3. *The upper bound of total expected regret of the improved UCB algorithm until trial T is*

$$\sum_{i\in A,\Delta_i>\lambda}\left(\Delta_i+\frac{32\log(T\Delta_i^2)}{\Delta_i}+\frac{96}{\Delta_i}\right)+\sum_{i\in A:0<\Delta_i\leq\lambda}\frac{64}{\lambda}+\max_{i\in A:\Delta_i\leq\lambda}\Delta_i T \tag{5}$$

for all $\lambda\geq\sqrt{\frac{e}{T}}$.

Proof of Theorem 3: For each sub-optimal arm i, we denote $m_i:=\min\{m|\tilde{\Delta}_m<\frac{\Delta_i}{2}\}$ as the *first* round in which $\tilde{\Delta}_m<\frac{\Delta_i}{2}$. Since it is the *first* round and $\tilde{\Delta}_m$ halves per round, we have

$$\tilde{\Delta}_{m_i+1}=\frac{\tilde{\Delta}_{m_i}}{2}=2^{-m_i-1}<\frac{\Delta_i}{4}\leq\tilde{\Delta}_{m_i}=2^{-m_i}<\frac{\Delta_i}{2}. \tag{6}$$

Let $A'=\{i\in A|\Delta_i>0\}$ and $A''=\{i\in A|\Delta_i>\lambda\}$ for some fixed $\lambda\geq\sqrt{\frac{e}{T}}$. Now we can analyze the regret in the following cases:

Case (i)**.** *Some sub-optimal arm $i\in A''$ is not eliminated in round m_i (or before) with an optimal arm $i^*\in B_{m_i}$.*

Consider an arbitrary sub-optimal arm $i\in A''$. At round m_i, if

$$\hat{\mu}_i\leq\mu_i+\sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}, \tag{7}$$

and

$$\hat{\mu}_{i^*}\geq\mu_{i^*}-\sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}, \tag{8}$$

both hold, then under the assumption $i^*,i\in B_{m_i}$, arm i will be eliminated in round m_i as shown in the following:

By (3) and (6), we derived that

$$\sqrt{\frac{\log\left(T\tilde{\Delta}_{m_i}^2\right)}{2n_{m_i}}}\leq\frac{\tilde{\Delta}_{m_i}}{2}=\tilde{\Delta}_{m_i+1}<\frac{\Delta_i}{4}. \tag{9}$$

Hence, in the arm elimination phase of round m_i,

$$\begin{aligned}\hat{\mu}_i+\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}}&\overset{(a)}{\leq}\mu_i+2\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}}\\&\overset{(b)}{<}\mu_i+\Delta_i-2\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}}\\&=\mu_{i^*}-2\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}}\\&\overset{(d)}{\leq}\hat{\mu}_{i^*}-\sqrt{\frac{\log(T\tilde{\Delta}_{m_i}^2)}{2n_{m_i}}},\end{aligned}$$

where (a) is (7), (b) is (9), and (d) is (8). Then, by (4), arm i is eliminated in round m_i as claimed. Now, by Chernoff-Hoeffding bounds [12], for any round m,

$$\mathbb{P}\left\{\underbrace{\hat{\mu}_i>\mu_i+\sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}}_{\text{opposite of (7)}}\right\}\leq\exp\left(-\frac{2n_m^2\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}{n_m}\right)=\frac{1}{T\tilde{\Delta}_m^2}, \tag{10}$$

and

$$\mathbb{P}\left\{\underbrace{\hat{\mu}_{i^*} < \mu_{i^*} - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}}_{\text{opposite of (8)}}\right\} \leq \exp\left(-\frac{2n_m^2 \frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}{n_m}\right) = \frac{1}{T\tilde{\Delta}_m^2}. \tag{11}$$

That is, the probability that a sub-optimal arm i is *not* eliminated in round m_i (or before) is bounded by $\frac{2}{T\tilde{\Delta}_{m_i}^2}$. At last, we obtain the contribution of **Case** (i) to the regret bound by summing over all arms in A' and bounding the regret for each arm i trivially by $T\Delta_i$ as

$$\sum_{i\in A''} \frac{2\Delta_i}{\tilde{\Delta}_{m_i}^2} \leq \sum_{i\in A''} \frac{8}{\tilde{\Delta}_{m_i}} \leq \sum_{i\in A''} \frac{32}{\Delta_i}, \tag{12}$$

where all inequality is due to (6).

Case (ii)**.** *For each sub-optimal arm i: either i is eliminated in round m_i (or before) or there is no optimal arm i^* in B_{m_i}.*

First consider the case that i is eliminated in round m_i. Since i is eliminated in round m_i (or before), according to the algorithm, arm i is pulled no more than

$$n_{m_i} = \left\lceil \frac{2\log(T\tilde{\Delta}_{m_i}^2)}{\tilde{\Delta}_{m_i}^2} \right\rceil \leq \left\lceil \frac{32\log(T\frac{\Delta_i^2}{4})}{\Delta_i^2} \right\rceil \tag{13}$$

times, where the inequality is due to (6). Thus, in this case, the contribution to the regret bound is

$$\sum_{i\in A''} \Delta_i \left\lceil \frac{32\log(T\frac{\Delta_i^2}{4})}{\Delta_i^2} \right\rceil < \sum_{i\in A''} \left(\Delta_i + \frac{32\log(T\Delta_i^2)}{\Delta_i}\right) \tag{14}$$

Next, we consider the other case that the last remaining optimal arm i^* is eliminated by some suboptimal arm $i \in A'$ in some round m_*. First note that if (7) and (8) hold in round m_*, then we have

$$\mu_i \overset{(a)}{\geq} \hat{\mu}_i - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}} \overset{(b)}{>} \hat{\mu}_* + \sqrt{\frac{\log(T\tilde{\Delta}_{m_*}^2)}{2n_{m_*}}} \overset{(c)}{\geq} \mu^*, \tag{15}$$

where (a) is (7), (b) is due to the fact that i^* is eliminated by i in round m_* (i.e. (4)), and (c) is (8). Then, we have a contradiction that $\mu_i > \mu^*$ for a sub-optimal arm i. Thus, the probability that i^* is eliminated by a sub-optimal arm i in round m_* is upper bounded by $\frac{2}{T\tilde{\Delta}_{m_*}^2}$, according to (10) and (11).

Now, i^* is eliminated by arm i in round m_*, then $i^* \in B_{m_j}$ for all j with $m_j < m_*$. Hence, by the assumption of **Case** (ii), all arms j with $m_j < m_*$ were eliminated in round m_j (or before). Therefore i^* can only be eliminated in round m_* by an arm i with $m_i \geq m_*$. Additionally, the maximum regret per trial after m_* is the maximum Δ_j among the remaining arms j with $m_j \geq m_*$. Let $m_\lambda := \min\{m|\tilde{\Delta}_m < \frac{\lambda}{2}\}$, then we have

$$\max_{j\in A''} m_j \leq m_\lambda \leq \min_{i\in A'\setminus A''} m_i, \tag{16}$$

according to the definition of A' and A''. Thus, for this case, the contribution of this case is upper bounded by

$$\sum_{m_*=0}^{\max_{j\in A''} m_j} \sum_{i\in A': m_i \geq m_*} \left(\frac{2}{T\tilde{\Delta}_{m_*}^2} T \max_{j\in A': m_j \geq m_*} \Delta_j\right)$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sum_{m_*=0}^{m_\lambda} \sum_{i\in A':m_i\geq m_*} \left(\frac{2}{\tilde{\Delta}_{m_*}^2} 4\tilde{\Delta}_{m_*}\right) \\
&= \sum_{m_*=0}^{m_\lambda} \left(\sum_{i\in A'\setminus A'':m_i\geq m_*} \frac{8}{\tilde{\Delta}_{m_*}} + \sum_{i\in A'':m_i\geq m_*} \frac{8}{\tilde{\Delta}_{m_*}} \right) \\
&= \sum_{m_*=0}^{m_\lambda} \left(\sum_{i\in A'\setminus A''} 8\cdot 2^{m_*} + \sum_{i\in A'':m_i\geq m_*} 8\cdot 2^{m_*} \right) \\
&= \sum_{i\in A'\setminus A''} \sum_{m_*=0}^{m_\lambda} 8\cdot 2^{m_*} + \sum_{i\in A''} \sum_{m_*=0}^{m_i} 8\cdot 2^{m_*} \\
&< \sum_{i\in A'\setminus A''} 8\cdot 2^{m_\lambda+1} + \sum_{i\in A''} 8\cdot 2^{m_i+1} \\
&\overset{(b)}{\leq} \sum_{i\in A'\setminus A''} 8\cdot \frac{8}{\lambda} + \sum_{i\in A''} 8\cdot \frac{8}{\Delta_i} \\
&= \sum_{i\in A'\setminus A''} \frac{64}{\lambda} + \sum_{i\in A''} \frac{64}{\Delta_i},
\end{aligned}
$$

where (a) is due to (6) and (16), and (b) is due to (6).

Finally, we can sum up the all contributions of considered cases and those suboptimal arms not in A'' to get the claimed upper bound.

6 Upper Confidence Bound with Variance-Aware (UCBV)

In [2], the authors proposed a new algorithm that considered variance of each arm. The algorithm, as shown in Algorithm 5, is similar to UCB1: it simply chooses an arm in each trial that maximizes a certain value. Nevertheless, with variance-aware, this algorithm uses Bernstein's inequality [14] instead of Chernoff-Hoeffding bounds [12] to bound the regret and obtain a better bound than UCB1.

Algorithm 4: UCBV algorithm

Input: A stochastic MAB problem as defined in section 2.

for $t \coloneqq 1$ **to** T **do**

Arm selection:

If $t \leq K$, select arm $i_t = t$. (Initialization).

If $t > K$, select arm $i_t \in \arg\max_{i\in A} B_{i,s,t}$, where

$$B_{i,s,t} \coloneqq \hat{\mu}_i^s + \sqrt{\frac{2\hat{v}_i^s \mathcal{E}_t}{s}} + c\frac{3\mathcal{E}_t}{s}, \tag{17}$$

and $c \geq 0, \hat{v}_i^s$ is the empirical variance of arm i after being pulled s times, $\mathcal{E}_t$ is a non-negative, non-decreasing function of t.

end

Theorem 4. *The upper bound of total expected regret of the UCBV algorithm until trial T is*

$$\sum_{i:\Delta_i>0} \left\{ 1 + 8\max\{c,1\}\left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i}\right)\mathcal{E}_T + Te^{-\mathcal{E}_T}\left(\frac{24\sigma_i^2}{\Delta_i^2} + \frac{4}{\Delta_i}\right) + \sum_{t=16\mathcal{E}_T}^{T} \beta(\min\{c,1\}\mathcal{E}_t, t) \right\} \Delta_i, \tag{18}$$

where $\beta(x,t) = 3\inf_{1<\alpha\leq 3} \min\{\frac{\log t}{\log \alpha}, t\} e^{-x/\alpha}$, and $\mathcal{E}_t, c$ is the same as in (17).

Proof of Theorem 18: We first show several lemma before the main proof.

Lemma 5. *In trial t, for any arm i, and $x > 0$, wit probability at least $1 - \beta(x, t)$*

$$|\hat{\mu}_i^s - \mu_i| \leq \sqrt{\frac{2\hat{v}_i^s x}{s}} + \frac{3x}{s} \tag{19}$$

hold for all $s \in [t]$

Lemma 6. *Consider after K plays, each arms has been pulled once. Let arm i be fixed. For any $\tau \in \mathbb{R}$, and any integer $u > 1$, we have*

$$\mathbb{E}[N_{i,t-1}] \leq u + \sum_{t=u+K-1}^{T} \sum_{s=u}^{t-1} \mathbb{P}\{B_{i,s,t} > \tau\} + \sum_{t=u+K-1}^{T} \mathbb{P}\{\exists s : 1 \leq s \leq t-1 \ s.t. \ B_{i^*,s,t} \leq \tau\} \tag{20}$$

The proof of these lemmas can be found in [2]. Now, let's start to proof theorem 18. According to (1), it suffices to bound $\mathbb{E}[N_{i,t-1}]$, where i is a sub-optimal arm. Thus, we fixed i. Let $\mathcal{E}'_n = \max\{c, 1\}\mathcal{E}_n$. Using (20) with $u = \left\lceil 8(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i})\mathcal{E}'_n \right\rceil$ and $\tau = \mu^*$, we guarantees that $\forall u \leq s < t$, and $t \geq 2$,

$$\begin{aligned}
\sqrt{\frac{2(\sigma_i^2 + \Delta_i/2)\mathcal{E}_t}{N_{i,t-1}}} + 3c\frac{\mathcal{E}_t}{N_{i,t-1}} &\leq \sqrt{\frac{(2\sigma_i^2 + \Delta_i)\mathcal{E}'_T}{u}} + 3\frac{\mathcal{E}'_n}{u} \\
&\overset{(a)}{\leq} \sqrt{\frac{(2\sigma_i^2 + \Delta_i)\Delta_i^2}{8(\sigma_i^2 + 2\Delta_i)}} + \frac{3\Delta_i^2}{8(\sigma_i^2 + 2\Delta_i)} \\
&= \frac{\Delta_i}{2}\left(\sqrt{\frac{2\sigma_i^2 + \Delta_i}{2\sigma_i^2 + 4\Delta_i}} + \frac{3\Delta_i}{4\sigma_i^2 + 8\Delta_i}\right) \\
&\overset{(b)}{\leq} \frac{\Delta_i}{2},
\end{aligned} \tag{21}$$

where (a) is from the choice of u, and (b) is equivalent to $(x-1)^2 \geq 0$, for $x = \sqrt{\frac{2\sigma_i^2 + \Delta_i}{2\sigma_i^2 + 4\Delta_i}}$.

For any $s \geq u$, and $t \geq 2$, with (21), we have

$$\begin{aligned}
&\mathbb{P}\{B_{i,s,t} > \mu^*\} = \mathbb{P}\left\{\hat{\mu}_i^s + \sqrt{\frac{2\hat{v}_i^s \mathcal{E}_t}{s}} + 3c\frac{\mathcal{E}_t}{s} > \mu_i + \Delta_i\right\} \\
&\leq \mathbb{P}\left\{\hat{\mu}_i^s + \sqrt{\frac{2(\sigma_i^2 + \Delta_i/2)\mathcal{E}_t}{s}} + 3c\frac{\mathcal{E}_t}{s} > \mu_i + \Delta_i\right\} + \mathbb{P}\left\{\hat{v}_i^s \geq \sigma_i^2 + \frac{\Delta_i}{2}\right\} \\
&\leq \mathbb{P}\left\{\hat{\mu}_i^s - \mu_i > \Delta_i/2\right\} + \mathbb{P}\left\{\frac{\sum_{j=1}^{s}(\hat{\mu}_i^j - \mu_i)^2}{s} - \sigma_i^2 \geq \Delta_i/2\right\} \\
&\leq 2\exp\left(-\frac{s\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3}\right),
\end{aligned} \tag{22}$$

which in the last step we used Bernstein's inequality [14] twice. Now, summing up these probabilities, we obtain

$$\begin{aligned}
\sum_{s=u}^{t-1} \mathbb{P}\{B_{i,s,t} > \mu^*\} &\leq 2\sum_{s=u}^{\infty} \exp\left(\frac{-s\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3}\right) \\
&= 2\frac{\exp(\frac{-u\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3})}{1 - \exp(-\frac{\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3})}
\end{aligned}$$

$$\overset{(a)}{\leq} \left(\frac{24\sigma_i^2}{\Delta_i^2} + \frac{4}{\Delta_i}\right) \exp\left(\frac{u\Delta_i^2}{8\sigma_i^2 + 4\Delta_i/3}\right)$$
$$\leq \left(\frac{24\sigma_i^2}{\Delta_i^2} + \frac{4}{\Delta_i}\right) e^{\mathcal{E}'_T},$$

where (a) is due to $1 - e^{-x} \geq \frac{2x}{3}$, for $0 \leq x \leq \frac{3}{4}$. For another term in (20), we use (19) to bound it as $\beta(\mathcal{E}'_n, t)$. Summing all contributions according to (20), we have

$$\mathbb{E}[N_{i,t-1}] \leq u + Te^{-\mathcal{E}'_T}\left(\frac{24\sigma_i^2}{\Delta_i^2} + \frac{4}{\Delta_i}\right) + \sum_{t=u+K-1}^{T} \beta(\mathcal{E}'_T, t)$$
$$\leq 1 + 8\mathcal{E}'_T\left(\frac{\sigma_i^2}{\Delta_i^2} + \frac{2}{\Delta_i}\right) + Te^{-\mathcal{E}'_T}\left(\frac{24\sigma_i^2}{\Delta_i^2} + \frac{4}{\Delta_i}\right) + \sum_{t=u+1}^{T} \beta(\mathcal{E}'_T, t).$$

Finally, by assumption $u \geq 16\mathcal{E}_T$, and summing over all sub-optimal arm gives the claimed bound.

Corollary 7. *If $c = 1$, and $\mathcal{E}_t = \zeta \log t$ for $\zeta > 1$, then there exists a constant c_ζ depending only on ζ such that for $T \geq 2$*

$$\mathbb{E}[R_T] \leq c_\zeta \sum_{i:\Delta_i>0} \left(\frac{\sigma_i^2}{\Delta_i} + 2\right) \log T.$$

For example, for $\zeta = 1.2$, the result holds for $c_\zeta = 10$.

Proof of Corollary 7: The proof is quite technical, so we omit here. However, the proof can be found in [2].

7 Efficient-UCBV (EUCBV)

Proposed in 2018, Efficient-UCBV (EUCBV) is the most advanced UCB-variant currently. It incorporates the arm elimination strategy proposed in Improved-UCB [3], while taking into account the variance estimates to compute the arms' confidence interval, similar to [2].

The algorithm of EUCBV is shown in Algorithm 5, and theorem 8 establishes the regret of EUCBV.

Algorithm 5: Efficient-UCBV (EUCBV) algorithm

Input: A stochastic MAB problem as defined in section 2.

$m \coloneqq 0, \tilde{\Delta}_0 \coloneqq 1, M \coloneqq \lfloor \frac{1}{2}\log_2 \frac{T}{e} \rfloor, n_0 \coloneqq \left\lceil \frac{\log(\psi T \tilde{\Delta}_0^2)}{2\tilde{\Delta}_0} \right\rceil, N_0 \coloneqq Kn_0$, and $B_0 \coloneqq A$

Pull each arm once (initialization).

for $t \coloneqq K+1$ **to** T **do**

Arm selection:

Pull arm $i \in \arg\max_{j \in B_m} \left\{ \hat{\mu}_j + \sqrt{\frac{\rho(\hat{v}_j+2)\log(\psi T \tilde{\Delta}_m)}{4N_{i,t-1}}} \right\}$.

Arm elimination:

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

$$\hat{\mu}_i + \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T \tilde{\Delta}_m)}{4N_{i,t-1}}} < \max_{j \in B_m} \left\{ \hat{\mu}_j - \sqrt{\frac{\rho(\hat{v}_j+2)\log(\psi T \tilde{\Delta}_m)}{4N_{i,t-1}}} \right\}, \tag{23}$$

to obtain B_{m+1}.

Reset Parameters:

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

$\tilde{\Delta}_{m+1} \coloneqq \frac{\tilde{\Delta}_m}{2}$

$n_{m+1} \coloneqq \left\lceil \frac{\log(\psi T \tilde{\Delta}_{m+1}^2)}{2\tilde{\Delta}_{m+1}} \right\rceil$

$N_{m+1} \coloneqq t + |B_{m+1}| n_{m+1}$

$m \coloneqq m+1$

end

Stop if $|B_m| = 1$ and pull $i \in B_m$ till T is reached.

end

Theorem 8. *The upper bound of total expected regret of the EUCBV algorithm until trial T is*

$$\sum_{i \in A: \Delta_i > \lambda} \left(\frac{C_0 K^4}{T^{\frac{1}{4}}} + \Delta_i + \frac{320\sigma_i^2 \log(\frac{T\Delta_i^2}{K})}{\Delta_i} \right) + \sum_{i \in A: 0 < \Delta_i \leq \lambda} \frac{C_2 K^4}{T^{\frac{1}{4}}} + \max_{i \in A: 0 < \Delta_i \leq \lambda} \Delta_i T. \tag{24}$$

for all $\lambda \geq \sqrt{\frac{e}{T}}$, and C_0, C_2 are integer constants.

Proof of Theorem 8: Let $m_i = \min\{m | \sqrt{4\tilde{\Delta}_m} < \frac{\Delta_i}{4}\}, \psi = \frac{T}{K^2}, \rho = \frac{1}{2}, c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T \tilde{\Delta}_{m_i})}{4N_{i,t-1}}}, n_{m_i} = \left\lceil \frac{\log(\psi T \tilde{\Delta}_{m_i}^2)}{2\tilde{\Delta}_{m_i}} \right\rceil$.

We need to first show several technical lemmas before the main proof.

Lemma 9. *If $T \geq K^{2.4}$, and $m \leq \frac{1}{2}\log_2(\frac{T}{e})$, then,*

$$\frac{\rho m \log(2)}{\log(\psi T) - 2m\log(2)} \leq \frac{3}{2}. \tag{25}$$

Lemma 10. *If $T \geq K^{2.4}$, then,*

$$c_i < \frac{\Delta_i}{4}. \tag{26}$$

Lemma 11. *In the m_i-th round,*

$$\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \frac{2}{(\psi T \tilde{\Delta}_{m_i})^{\frac{3\rho}{2}}}. \tag{27}$$

Lemma 12. *In the m_i-th round,*

$$\mathbb{P}\{c^* > c_i\} \leq \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}}. \tag{28}$$

Lemma 13. *In the m_i-th round,*

$$\mathbb{P}\{N_{i,t-1} < n_{m_i}\} \leq \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}}. \tag{29}$$

Lemma 14. *For two integer constants c_1 and c_2, if $20c_1 \leq c_2$ then,*

$$c_1 \frac{4\sigma_i^2 + 4}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right) \leq c_2 \frac{\sigma_i^2}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right). \tag{30}$$

The proofs of these lemmas can be found in the appendix of [4]. Now, we can start to proof theorem 8. Let $A' = \{i \in A | \Delta_i > 0\}$, and $A'' = \{i \in A | \Delta_i \geq b\}$. Note that as rewards are bounded in $[0, 1]$, it implies that $0 \leq \sigma_i^2 \leq \frac{1}{4}$. We follow the steps of [3] to bound the regret under two cases.

Case (i)**.** *Some sub-optimal arm $i \in A''$ is not eliminated in round m_i (or before) with an optimal arm $i^* \in B_{m_i}$.*

Since c_i is the length of the confidence interval of arm i in round m_i, whenever $N_{i,t-1} \geq m_i \geq \frac{\log(\psi T \tilde{\Delta}_{m_i}^2)}{2\tilde{\Delta}_{m_i}}$ we have $c_i \leq \frac{\Delta_i}{4}$ by lemma 10. The sufficient conditions for arm i to be eliminated in round m_i is

$$(1)\hat{\mu}_i \leq \mu_i + c_i, (2)\hat{r}^* \geq r^* - c^*, (3)c_i \geq c^*, and (4)N_{i,t-1} \geq n_{m_i} \tag{31}$$

Suppose (31) holds, then we have

$$\begin{aligned}\hat{\mu}_i + c_i &\leq \mu_i + 2c_i = \mu_i + 4c_i - 2c_i \\ &< \mu_i + \Delta_i - 2c_i \leq \mu^* - 2c^* \leq \hat{\mu}^* - c^*,\end{aligned}$$

so that a sub-optimal arm $i \in A''$ gets eliminated. Hence, the probability of the complementary event of the four conditions in (31) yields a bound on the probability that arm i is *not* eliminated in round m_i. According to the proof of Lemma 1 in [2], we can show the bound is

$$\mathbb{P}(\hat{\mu}_i > \mu_i + c_i) \leq \mathbb{P}(\hat{\mu}_i > \mu_i + \bar{c}_i) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\tilde{\Delta}_{m_i}}\right), \tag{32}$$

where

$$\bar{c}_i = \sqrt{\frac{\rho(\sigma_i^2 + \sqrt{\tilde{\Delta}_{m_i}} + 2)\log(\psi T \tilde{\Delta}_{m_i})}{4n_{m_i}}}. \tag{33}$$

From lemma 11, we can show that

$$\mathbb{P}(\hat{\mu}_i > \mu_i + c_i) \leq \mathbb{P}(\hat{\mu}_i > \mu_i + \bar{c}_i) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\tilde{\Delta}_{m_i}}\right) \leq \frac{2}{(\psi T \tilde{\Delta}_{m_i})^{\frac{3\rho}{2}}}.$$

Similarly,

$$\mathbb{P}\{\hat{\mu}^* < \mu^* - c^*\} \leq \frac{2}{(\psi T \tilde{\Delta}_{m_i})^{\frac{3\rho}{2}}}.$$

Summing the above two terms, the probability that a sub-optimal arm $i \in A''$ is not eliminated on (or before) m_i-th round by the first two conditions in (31) is

$$\frac{4}{(\psi T \tilde{\Delta}_{m_i})^{\frac{3\rho}{2}}}. \tag{34}$$

Again, from lemma 12 and lemma 13, we bound the probability of the opposite event $c_i \geq c^*$ and $N_{i,t-1} \geq n_{m_i}$ by

$$\frac{182K^4}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}} + \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}} \leq \frac{364K^4}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}}. \tag{35}$$

Also, for (34), $\forall \tilde{\Delta}_{m_i} \in [\sqrt{\frac{e}{T}}, 1]$

$$\frac{4}{(\psi T \tilde{\Delta}_{m_i})^{\frac{3\rho}{2}}} = \frac{4}{(\frac{T^2}{K^2}\tilde{\Delta}_{m_i})^{\frac{3}{4}}} = \frac{4K^{\frac{3}{2}}}{(T^{\frac{3}{2}}\tilde{\Delta}_{m_i}^{\frac{1}{4}}\sqrt{\tilde{\Delta}_{m_i}})} \overset{(a)}{\leq} \frac{4K^{\frac{3}{2}}}{(T^{\frac{3}{2}-\frac{1}{8}}\sqrt{\tilde{\Delta}_{m_i}})} \overset{(b)}{\leq} \frac{4K^4}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}}, \tag{36}$$

where (a) follows from the identity $\tilde{\Delta}_{m_i}^{\frac{1}{4}} \geq (\frac{e}{T})^{\frac{1}{8}}$ as $\tilde{\Delta}_{m_i} \geq \sqrt{\frac{e}{T}}$, and (b) assumes $T \geq K^{2.4}$.

Summing over all arms in A'' and bounding the regret for (31) by (35) + (36) for each arm $i \in A''$ trivially by $T\Delta_i$, we obtain

$$\sum_{i\in A''}\left(\frac{4K^4T\Delta_i}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}}\right) + \sum_{i\in A''}\left(\frac{364K^4T\Delta_i}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_i}}}\right) \overset{(a)}{\leq} \sum_{i\in A''}\left(\frac{368K^4T\Delta_i}{T^{\frac{5}{4}}(\frac{\Delta_i^2}{4.16})^{\frac{1}{2}}}\right) \overset{(b)}{\leq} \sum_{i\in A''}\left(\frac{C_1K^4}{T^{\frac{1}{4}}}\right), \tag{37}$$

where (a) is due to $\sqrt{4\tilde{\Delta}_{m_i}} < \frac{\Delta_i}{4}$, and in (b), C_1 is a constant integer.

Case (ii)**.** *For each sub-optimal arm i: either i is eliminated in round m_i (or before) or there is no optimal arm i^* in B_{m_i}.*

First consider the case that i is eliminated in round m_i. Since i is eliminated in round m_i (or before), according to the algorithm, arm i is pulled no more than

$$N_{i,t-1} < \left\lceil \frac{\log(\psi T \tilde{\Delta}_{m_i}^2)}{2\tilde{\Delta}_{m_i}} \right\rceil.$$

Thus, the total contribution of arm i until round m_i is

$$\Delta_i\left\lceil \frac{\log(\psi T \tilde{\Delta}_{m_i}^2)}{2\tilde{\Delta}_{m_i}} \right\rceil \overset{(a)}{\leq} \Delta_i\left\lceil \frac{\log(\psi T(\frac{\Delta_i}{16\cdot 256})^4)}{2(\frac{\Delta_i}{4\sqrt{4}})^2} \right\rceil \leq \Delta_i\left(1 + \frac{32\log(\psi T(\frac{\Delta_i^4}{16384}))}{\Delta_i^2}\right) \leq \Delta_i\left(1 + \frac{32\log(\psi T\Delta_i^4)}{\Delta_i^2}\right),$$

where (a) is due to $\sqrt{4\tilde{\Delta}_{m_i}} < \frac{\Delta_i}{4}$. Now, summing over all arms in A'' gives

$$\begin{aligned}
&\sum_{i\in A''}\Delta_i\left(1 + \frac{32\log(\psi T\Delta_i^4)}{\Delta_i^2}\right) = \sum_{i\in A''}\left(\Delta_i + \frac{32\log(\psi T\Delta_i^4)}{\Delta_i}\right) \\
&= \sum_{i\in A''}\left(\Delta_i + \frac{64\log(\frac{T\Delta_i^2}{K})}{\Delta_i}\right) \\
&\overset{(a)}{\leq} \sum_{i\in A''}\left(\Delta_i + \frac{16(4\sigma_i^2+4)\log(\frac{T\Delta_i^2}{K})}{\Delta_i}\right)
\end{aligned}$$

$$\overset{(b)}{\leq} \sum_{i \in A''} \left(\Delta_i + \frac{320\sigma_i^2 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right),$$

where (a) is from $0 \leq \sigma_i^2 \leq \frac{1}{4}$, and (b) is from lemma 14.

Next, we consider the other case that the last remaining optimal arm i^* is eliminated by some suboptimal arm $i \in A'$ in some round m_*. Similar as the discussion in [3], if the conditions of **Case** (i) (i.e. (31)) holds, then the optimal arm i^* will not be eliminated in round m_*, or elase $r_i > r^*$. Again, as in [3], since i^* is eliminated in m_*, then any arm j with $m_j < m_*$ is eliminated according to the assumption in **Case** (ii). Therefore i^* can only be eliminated in round m_* by an arm i with $m_i \geq m_*$. Additionally, the maximum regret per trial after m_* is the maximum Δ_j among the remaining arms j with $m_j \geq m_*$. Let $m_\lambda = \min\{m | \sqrt{4\tilde{\Delta}_m} < \frac{\lambda}{4}\}$. Then, in this case, the regret is bounded by

$$\begin{aligned}
&\sum_{m_*=0}^{max_{j \in A''} m_j} \sum_{i \in A': m_i > m_*} \frac{368K^4}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_*}}} \cdot T \cdot \max_{j \in A': m_j \geq m_*} \Delta_j \\
&\leq \sum_{m_*=0}^{max_{j \in A''} m_j} \sum_{i \in A': m_i > m_*} \frac{368K^4\sqrt{4}}{T^{\frac{5}{4}}\sqrt{\tilde{\Delta}_{m_*}}} \cdot T \cdot 4\sqrt{\tilde{\Delta}_{m_*}} \\
&\overset{(a)}{\leq} \sum_{m_*=0}^{max_{j \in A''} m_j} \sum_{i \in A': m_i > m_*} \frac{C_2 K^4}{T^{\frac{1}{4}} \tilde{\Delta}_{m_*}^{\frac{1}{2}-\frac{1}{2}}} \\
&\leq \sum_{i \in A': m_i > m_*} \sum_{m_*=0}^{\min\{m_i, m_\lambda\}} \frac{C_2 K^4}{T^{\frac{1}{4}}} \\
&\leq \sum_{i \in A''} \frac{C_2 K^4}{T^{\frac{1}{4}}} + \sum_{i \in A' \setminus A''} \frac{C_2 K^4}{T^{\frac{1}{4}}}.
\end{aligned}$$

At (a), C_2 is an integer constant. Here, the derivation process is very similar to that of ([3]).

Finally, adding all cases up and with arms $i \in A' \setminus A''$, the total expected regret bound is given by

$$\sum_{i \in A: \Delta_i > \lambda} \left(\frac{C_0 K^4}{T^{\frac{1}{4}}} + \Delta_i + \frac{320\sigma_i^2 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right) + \sum_{i \in A: 0 < \Delta_i \leq \lambda} \frac{C_2 K^4}{T^{\frac{1}{4}}} + \max_{i \in A: 0 < \Delta_i \leq \lambda} \Delta_i T,$$

where $C_2, C_0 = C_1 + C_2$ are integer constants.

8 Regret Lower Bound and Optimality of KL-UCB

We say algorithm A is *consistent* if for any sub-optimal arm $i : \Delta_i > 0$ and and $a > 0$

$$\mathbb{E}[N_{i,T}] = o(T^a)$$

i.e. The algorithm eventually samples any sub-optimal arm a vanishing fraction of times.
A classical result by Lai & Robbins [7] states that if algorithm A is consistent and reward $X_i \sim \text{Bernoulli}(\mu_i), \forall i$ then

$$\liminf_{T \to \infty} \frac{\mathbb{E}[R_T[A]]}{\log T} \geq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{KL(\mu_i, \mu^*)} \tag{38}$$

We note that Garivier & Cappe [8] gave the KL-UCB algorithm that makes use upper confidence bounds derived from KL-divergence based concentration inequalities rather than Hoeffding's inequality, and that matches the constant factor in the lower bound above.
Below we show the KL-UCB algorithm and it's regret upper bound. Denote $d(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ the KL-divergence of Bernoulli(p) and Bernoulli(q).

Algorithm 6: KL-UCB algorithm

Input: A stochastic MAB problem as defined in section 2.
for $t := 1$ **to** T **do**
  Arm selection:
  If $t \leq K$, select arm $i_t = t$. (Initialization).
  If $t > K$, select arm $i_t \in \arg\max_{i\in A} \max_{q\in[0,1]}\{N_{i,t-1} \times d(\hat{\mu}_i^{t-1}, q) \leq \log t + 3\log\log t\}$.
end

We have the following theorem:

Theorem 15. *For any sub-optimal arm $i : \Delta_i > 0$ and any $\epsilon > 0$:*

$$\mathbb{E}[N_{i,T}] \leq \frac{\log T}{d(\mu_i, \mu^*)}(1+\epsilon) + C_1 \log\log T + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}$$

where C_1 is a positive constant and $C_2(\epsilon) = O(\epsilon^{-2}), \beta(\epsilon) = O(\epsilon^2)$ are both positive function of ϵ.

We omit the proof of the theorem since it is tedious, it can be found in [8] with some change of notation.
Remark:
Note that Theorem 15 implies

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_{i,T}]}{\log T} \leq \frac{1}{d(\mu_i, \mu^*)}$$

which means that if rewards are Bernoulli, KL-UCB algorithm fits the optimal bound. Also, by Pinsker's Inequality, $d(\mu_i, \mu^*) > 2(\mu_i - \mu^*)^2 = 2\Delta_i^2$, we can see that KL-UCB is strictly better than UCB1.
From table 1 we can conclude that all algorithms mentioned in this survey are consistent. One may notice that consistent is a good property, so maybe we should have the lower bound 38 for any policy. But [15] shows that for a weaker condition (called α-*consistent* in their work) we can get a even smaller lower bound which is asymptotically optimal. The result is counterintuitive and we can't understand the proof. So it will be left as our future exploration.

9 Conclusions

In this paper, we showcase several UCB-based algorithms that use different strategies to solve the stochastic MAB problem, and thus achieve different regret upper bound. We also demonstrate the regret lower bound for any consistent algorithm with Bernoulli rewards and state the optimality of KL-UCB in this special setting[1].

References

[1] Auer P.; and Cesa-Bianchi N. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, pages 235–256, 2002.

[2] Audibert J.-Y.; Munos R.; and Szepesvari C. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, pages 1876–1902, 2009.

[3] P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, page 55–65, 2010.

[4] Mukherjee S.; Naveen K.P.; Sudarsanam N.; and Ravindran B. Efficient-ucbv: An almost optimal algorithm using variance estimates. *AAAI Conference on Artificial Intelligence*, 2018.

[5] Audibert J.-Y.; and Bubeck S. Minimax policies for adversarial and stochastic bandits. *COLT*, page 217–226, 2009.

[6] Lattimore T. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint*, 2015.

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

[8] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

[9] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, pages 285–294, 1933.

[10] P. Auer; N. Cesa-Bianchi; Y. Freund; and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, page 48–77, 2002.

[11] J. Gittins; K. Glazebrook; and R. Weber. Multi-armed bandit allocation indices. *John Wiley and Sons*, 2011.

[12] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.

[13] S. Mannor E. Even-Dar and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, page 1079–1105, 2006.

[14] S.N.Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sec. Math.*, 1924.

[15] Antoine Salomon, Jean-Yves Audibert, and Issam El Alaoui. Regret lower bounds and extended upper confidence bounds policies in stochastic multi-armed bandit problem. *arXiv preprint arXiv:1112.3827*, 2011.