Stochastic One-Sided Full-Information Bandit

Haoyu Zhao[1] (✉) and Wei Chen[2]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China, zhaohy16@mails.tsinghua.edu.cn
[2] Microsoft Research, Beijing, China, weic@microsoft.com

Abstract. In this paper, we study the stochastic version of the one-sided full information bandit problem, where we have K arms $[K] = \{1, 2, \ldots, K\}$, and playing arm i would gain reward from an unknown distribution for arm i while obtaining reward feedback for all arms $j \geq i$. One-sided full information bandit can model the online repeated second-price auctions, where the auctioneer could select the reserved price in each round and the bidders only reveal their bids when their bids are higher than the reserved price. In this paper, we present an elimination-based algorithm to solve the problem. Our elimination based algorithm achieves distribution independent regret upper bound $O(\sqrt{T \cdot \log(TK)})$, and distribution dependent bound $O((\log T + \log K) f(\Delta))$, where T is the time horizon, Δ is a vector of gaps between the mean reward of arms and the mean reward of the best arm, and $f(\Delta)$ is a formula depending on the gap vector that we will specify in detail. Our algorithm has the best theoretical regret upper bound so far. We also validate our algorithm empirically against other possible alternatives.

Keywords: Online Learning · Multi-armed Bandit

1 Introduction

Stochastic multi-armed bandit (MAB) has been extensively studied in machine learning and sequential decision making. The most simple version of this problem consists of K arms, where each arm has an unknown distribution of the reward. The task is to sequentially select one arm at each round so that the total expected reward is as high as possible. In each round, we will gain the reward and only observe the reward of the arm we choose. The trade-off between exploration and exploitation appears extensively in the MAB problem: On the one hand, one might try to play an arm which is played less to explore whether it is good, and on the other hand, one might choose to play the arm with the largest average reward so far to cumulate reward. MAB algorithms are measured by their *regret*, which is the difference in expected cumulative reward between the algorithm and the optimal algorithm that always chooses the best arm.

A variant of the stochastic MAB problem is the *one-sided full-information bandit*, where there is a set of arms $1, 2, \ldots, K$ and at round t we choose arm I_t, we will gain the reward of I_t at time t and observe the rewards of all arms $i \geq I_t$

at time t (Section 2). The adversarial version of the one-sided full-information bandit is first introduced in [7], and in this paper, we study it stochastic version.

One-sided full-information bandit can find applications such as in online auction. Consider for example the second-price auction with a reserve price. In each round, the seller (or auctioneer) sets a reserve price from a finite set of reserve price choices. Each buyer (or bidder) draws a value from its valuation distribution (unknown to the seller), and only submits her value as the bid when her value is at least as high as the reserve price. The seller would observe these bids, give the item to the highest bidder and collect the second highest bid price (including the reserve price) as its reward from the highest bidder. In this case, we can treat each reserve price as an arm. In each round t after the seller announces the reserve price r_t, she will see all bids higher than r_t, and thus she would know the reward she could collect for all reserve prices higher than or equal to r_t, which corresponds to the case of one-sided full-information feedback.[1]

In this paper, we present an elimination-based algorithm for the stochastic one-sided full-information bandit and prove the distribution-independent bound as $O(\sqrt{T(\log T + \log K)})$ and the distribution-dependent bound as $O((\log T + \log K)f(\Delta))$, where T is the time horizon, Δ is a vector of gaps between the mean reward of arms and the mean reward of the best arm, and $f(\Delta)$ is a formula depending on the gap vector that we will specify in Theorem 2 (Section 3). We also adopt an existing analysis to show a distribution-independent regret lower bound of $\Omega(\sqrt{T \log K})$ for this case (Section 4), which indicates that our algorithm achieves almost matching upper bound. We conduct numerical experiments to show that our algorithm significantly outperforms an existing algorithm designed for the adversarial case (Section 5). The empirical results also indicate that a UCB variant has better empirical performance, but it so far has no tight theoretical analysis, and thus our elimination-based algorithm is still the one with the best theoretical guarantee.

Due to space constraint, some proofs are moved to a supplementary material submitted together with the main paper.

1.1 Related Work

Multi-armed bandit: Multi-armed bandit (MAB) is originally introduced by Robbins [9], and has been extensively studied in the literature (c.f. [4, 5]). MAB could be either stochastic, where the rewards of arms are drawn from unknown distributions, or adversarial, where the rewards of arms are determined by an adversary. Our study in this paper belongs to the stochastic MAB category. The classical MAB algorithm includes UCB [2] and Thompson sampling [10] for the stochastic setting and EXP3 [3] for the adversarial setting.

Multi-armed bandit with graph feedback structure: One-sided full-information bandit can be viewed as a special case of the MAB problem with

[1] Note that the second-price auction is truthful in a single round, but in multi-rounds, it may not be truthful since the bidders may want to lower their bids first so that the seller would learn a lower reserve price. The truthfulness is not the main concern of this paper and its discussion is beyond the scope of this paper.

graph feedback structure. The arm feedback structure can be represented as a graph (undirected or directed, with or without self-loops), where vertices are arms, and when an arm is played, the rewards of all its neighbors (or out-neighbors) can be observed. The one-sided full-information bandit corresponds to a feedback graph with directed edges pointing from arm i to arm j for all $i \leq j$. The first paper that introduces MAB with graph feedback is [8]. The authors of this paper use the independent number and the clique-partition number to derive the upper and lower bound for the regret. The main results of [8] is the upper and lower bound for the regret for undirected graph feedback MAB problem. Although the bound is tight in the undirected case, there is a gap between the regret upper and lower bounds for directed graphs. When translated to our one-sided full information setting, their regret upper bound is $\tilde{O}(\sqrt{KT})$ but the lower bound is $\tilde{\Omega}(\sqrt{T})$, which are not as tight as we provide in this paper in both upper and lower bounds. In [1], the authors consider the adversarial MAB with general directed feedback graph and close the gap up to some logarithmic factors. However, when applying their results to the one-sided full-information bandit setting, their upper and lower bounds are all worse than ours by a logarithmic factor. Moreover, we provide distribution-dependent bound that only works for the stochastic setting. One-sided full-information bandit is originally proposed in [7], which studies the adversarial setting and proposes a variant of EXP3 algorithm EXP3-RTB to solve this problem in the adversarial setting. Their work focuses on the more general bandit on metric space, and ignores the difference in the logarithmic factors. Stochastic MAB with undirected graph feedback is studied in [6], which proposes a variant of UCB algorithm UCB-N that essentially acts as UCB but updates all observed arms instead of only the played arm in each round. The authors derive a regret upper bound based on the cliques in the feedback graph, but in the one-sided full-information setting the cliques are reduced to singletons and their regret result is reduced to the classical UCB, which is significantly worse than the regret of our algorithm. We include UCB-N in our experiments, which demonstrate good performance of UCB-N, but we cannot provide a better theoretical regret bound for it, and this task is left as a future work item.

2 Model

In this section, we specify a multi-armed bandit model called 'one-sided full information bandit', which is highly related with the online auction problem. Suppose that there are K arms $[K] = \{1, 2, \ldots, K\}$ in total. Each time we play the arm I_t at round t, we will observe the value of arm i, denoted as $X_i^{(t)}$, for all $i \geq I_t$. We study this problem under the stochastic settings, i.e. in each round t, the realized value $X_i^{(t)}$ is drawn from a distribution ν_i, and $X_i^{(t)}$ is independent to $X_i^{(t')}$, for all $t' < t$. The formal definition of the bandit model is given as follow.

Definition 1 (One-sided Full Information Bandit). *There is a set of arms $\{1, 2, \ldots, K\}$, and for each arm $i \in [K]$, it corresponds to an unknown dis-*

tribution ν_i with support $[0,1]$, where ν_i is the marginal distribution of ν with support $[0,1]^K$. In each round t, the environment draws a reward vector $X^{(t)} = (X_1^{(t)}, \ldots, X_K^{(t)})$, where $X^{(t)}$ is drawn from distribution ν. The player then chooses an arm I_t to play, gains the reward $X_{I_t}^{(t)}$ and observes the reward of arms $I_t, I_t+1, \ldots, K$, i.e. observes $X_i^{(t)}, \forall i \geq I_t$.

Remark 1. In the definition, we explicitly give the joint distribution ν to describe the value distribution of arms, and denote arm i's reward distribution ν_i as the marginal distribution of ν. This is to emphasize the fact that the distributions corresponding to different arms can be correlated.

The performance of the multi-armed bandit algorithm is measured by regret. In the stochastic bandit scenario, people will use the pseudo-regret to measure the performance more often. The pseudo regret is defined as follow,

Definition 2 (Pseudo-regret). *Let I_t denote the arm that is chosen by algorithm $\mathcal{A}$ to play at round t, then the pseudo-regret of the algorithm $\mathcal{A}$ for T rounds is defined as $\mathbb{E}[\sum_{i=t}^{T}(X_{i^*}^{(t)} - X_{I_t}^{(t)})]$, where i^* denotes the best arm in expectation, i.e. $\mathbb{E}[X_{i^*}^{(t)}] \geq \mathbb{E}[X_i^{(t)}]$ for all $i \in [K]$.*

In this paper, we only consider pseudo-regret, and henceforth, for convenience, we simply use the term regret to refer to pseudo-regret in the remaining text. For convenience, we will use $\mu_i = \mathbb{E}[X_i^{(t)}]$ to denote the mean of the reward of arm i, and μ_{i^*} to denote the mean of the best arm. We will also use $\Delta_i = \mu_{i^*} - \mu_i$ to denote the difference of the mean between arm i and the best arm i^*.

3 Algorithm and Regret Analysis

3.1 Elimination Based Algorithm

In this section, we present an elimination-based algorithm to tackle the one-sided full information stochastic bandit problem. We first show an algorithm with known time horizon T. Our algorithm can be generally described as: We maintain a set of arms S_t during the execution of the algorithm. At each round, we will play the arm that has the smallest index in S_t, i.e. $I_t \leftarrow \min_{i \in S_t} i$. At first, $S_1 = [K]$ is the set of all arms, and we will play arm 1 in the first round. At each time t we observe the rewards for the arms $I_t, I_{t+1}, \ldots, K$, update the empirical mean of each arm and update the set S_t into S_{t+1}. At each round t, we will delete the arms in S_t whose empirical means are much smaller than the best empirical mean in S_t. More specifically, we have

$$m_t = \operatorname*{argmax}_{i \in S_t} \hat{\mu}_{i,t},$$

where $\hat{\mu}_{i,t} = \frac{1}{t}\sum_{s=1}^{t} X_i^{(s)}$ is the empirical mean of arm i after t rounds, and

$$S_t = \{i \in S_{t-1} \mid \hat{\mu}_{m_{t-1},t-1} - \hat{\mu}_{i,t-1} \leq 2\rho_t\},$$

Algorithm 1 ELIM: Elimination-based algorithm with known time horizon T

Input: Time horizon T.
1: $S_0 \leftarrow \{1, 2, \ldots, K\}$.
2: $\forall i, \hat{\mu}_{i,0} = 0$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: $\quad \rho_t \leftarrow \sqrt{\frac{\ln(KT^2)}{2(t-1)}}$.(The confidence radius ρ_1 at time $t = 1$ is ∞).
5: $\quad m_{t-1} \leftarrow \text{argmax}_{i \in S_{t-1}} \hat{\mu}_{i,t-1}$.
6: $\quad S_t \leftarrow \{i \in S_{t-1} \mid \hat{\mu}_{m_{t-1},t-1} - \hat{\mu}_{i,t-1} \leq 2\rho_t\}$.
7: $\quad$ Play the arm j, where $j \leftarrow \min_{i \in S_t} i$.
8: $\quad$ Observe the reward $X_i^{(t)}, \forall i \geq j$.
9: $\quad \forall i \in S_t, \hat{\mu}_{i,t} \leftarrow \hat{\mu}_{i,t-1} \cdot \frac{t-1}{t} + X_i^{(t)} \cdot \frac{1}{t}$.
10: **end for**

where ρ_t is the *confidence radius* and $\rho_t = \sqrt{\frac{\ln(KT^2)}{2(t-1)}}$(The confidence radius ρ_1 at around $t = 1$ is ∞). Our whole algorithm is shown in Algorithm 1.

We will show that our algorithm has distribution-independent regret bounded $O(\sqrt{T(\ln K + \ln T)})$, where the best regret bound for one-sided full information bandit till now is $O(\sqrt{T \ln K \ln T})$, which is implied in [7]. Besides the distribution-independent bound, we also give a distribution-dependent bound. The following two theorems show our results, and their proofs will be provided in the next section.

Theorem 1. (Distribution independent regret bound) *Given the time horizon T, the regret of Algorithm 1 is bounded by $4\sqrt{2T \ln(KT^2)} + 3$.*

Theorem 2 (Distribution dependent regret bound). *Let $\{\Delta_{(i)}\}$ be a permutation of $\{\Delta_i \mid i \leq i^*\}$, such that $\Delta_{(1)} \geq \Delta_{(2)} \geq \cdots \geq \Delta_{(i*)} = 0$, and $C = 8 \ln(KT^2)$. Given time horizon T, the regret of Algorithm 1 is bounded by*

$$\Delta_{(1)} + \frac{C}{\Delta_{(1)}} + C \sum_{i=2}^{i^*-1} \left(\frac{1}{\Delta_{(i)}^2} - \frac{1}{\Delta_{(i-1)}^2} \right) \Delta_{(i)} + 2. \tag{1}$$

Note that the standard UCB algorithm will lead to $O(\sum_{i \in [K], \Delta_i > 0} \frac{1}{\Delta_i} \ln T) = O(\sum_{i=1}^{K-1} \frac{1}{\Delta_{(i)}} \ln T)$ distribution dependent regret. In Eq. (1), if we ignore the term $-\frac{1}{\Delta_{(i-1)}^2}$ in the summation, we could obtain the same order regret upper bound. Thus, the regret obtained above is typically better than the UCB regret. To see more clearly the difference, consider the case when the best arm i^* has mean $\mu_{i^*} = \frac{1}{2} + \varepsilon$ and all other arms $i \neq i^*$ have mean $\mu_i = \frac{1}{2}$, the original UCB will lead to $\frac{K-1}{\varepsilon} \ln T$ regret bound, and our algorithm will lead to $\frac{8 \ln(KT^2)}{\varepsilon} + 2 + \varepsilon$ regret bound. Also notice that in the distribution dependent bound, we only add up to i^*, which means that the arms which have indices larger than i^* will not contribute explicitly to the regret upper bound. This directly shows that the location of the best arm matters in our algorithm for one-sided MAB model.

Remark 2. Although the arms with indices larger than that of the best arm do not contribute explicitly to the regret bound, they do contribute to the constant 2 in Eq.(1) of Theorem 2. The contribution comes from a low probability case, which is shown in the proof in the next section.

In Algorithm 1, we assume that we know the time horizon T. Now, we apply the standard 'doubling trick' to get an algorithm with unknown time horizon T, which is shown in Algorithm 2. The distribution independent regret bound is given in Theorem 3.

Algorithm 2 Algorithm with unknown time horizon

1: **for** $i = 0, 1, \ldots$ **do**
2: In time horizon $2^i, 2^i + 1, \ldots, 2^{i+1} - 1$, run Algorithm 1 with time horizon 2^i.
3: **end for**

Theorem 3. *The regret of Algorithm 2 is bounded by* $20\sqrt{T\ln(KT^2)}+3\log_2 T+3$.

3.2 Proof of Theorem 1

Because we want to observe as many arms as possible, we would like to choose an arm with a small index(a small position). In this way, our algorithm maintains a set of arms S_t in each round t, which is the set of arms that are possible to be the best arm. We could let $S_t = [K]$ for each round, then this will lead to large regret, so we would like all arms in S_t have means 'close' to the mean of the best arm, and the best arm i^* is in the set S_t. In this way, we will define "a procedure is nice at round t" in Definition 4 to describe the event that the best arm is in S_t and all of the arms in S_t have means close to that of the best arm. Then we will show in Lemma 2 that the procedure is nice at all rounds $t \leq T$ with high probability. Finally, we will use this lemma to prove Theorem 1. To begin with, we have the following definition and a simple lemma.

Definition 3. *We call the sampling is nice at the beginning of round t if* $|\hat{\mu}_{i,t-1} - \mu_i| < \rho_t, \forall i \in S_{t-1}$, *where* $\rho_t = \sqrt{\frac{\ln(KT^2)}{2(t-1)}}, \forall t \geq 2$ *and* $\rho_1 = \infty$. *Let* $\mathcal{N}_t^s$ *denote this event.*

Lemma 1. *For each round* $t \geq 1$, $Pr\{\neg\mathcal{N}_t^s\} \leq \frac{2}{T^2}$.

The proof of this lemma is simple with an application of the Hoeffding's Inequality followed by a union bound. For more detail, please see Appendix. Then, we have the definition for "procedure is nice at round t" and the main lemma that shows that the procedure is nice happens uniformly at all rounds with high probability. The formal definition is shown in Definition 4 and the lemma is formally stated in Lemma 2.

Definition 4. *We say that the procedure is nice during the algorithm at round t if both of the following are satisfied,*

1. $i^* \in S_t$, *where* $i^* = \arg\max_{i\in[K]} \mu_i$.
2. $\forall i \in S_t, \mu_{i^*} - \mu_i \le 4\rho_t$.

Let $\mathcal{N}_t^p$ denote this event.

Lemma 2. *Let $\mathcal{M}_t = \bigcap_{s=1}^t \mathcal{N}_s^p$, then*

$$\forall t \in [T], Pr\{\neg\mathcal{M}_t\} \le \frac{2}{T}.$$

Proof. We partition the event $\neg\mathcal{M}_t$ into disjoint events, we have

$$\neg\mathcal{M}_t = \neg\mathcal{N}_1^p \cup (\mathcal{M}_1 \cap \neg\mathcal{N}_2^p) \cup \cdots \cup (\mathcal{M}_{t-1} \cap \neg\mathcal{N}_t^p).$$

Note that $\neg\mathcal{M}_t$ is the union of disjoint events, so we have

$$\Pr\{\neg\mathcal{M}_t\} = \Pr\{\neg\mathcal{N}_1^p\} + \sum_{s=2}^{t} \Pr\{\mathcal{M}_{s-1} \cap \neg\mathcal{N}_s^p\}.$$

First, it is obvious that $\Pr\{\neg\mathcal{N}_1^p\} = 0$, since $\mathcal{N}_1^p$ will always happen, then we just need to bound $\Pr\{\mathcal{M}_{s-1} \cap \neg\mathcal{N}_s^p\}$ for each $2 \le s \le t$. We have

$$\begin{aligned}\Pr\{\mathcal{M}_{s-1} \cap \neg\mathcal{N}_s^p\} =& \Pr\left\{\left(\bigcap_{r=1}^{s-1} \mathcal{N}_r^p\right) \cap \neg\mathcal{N}_s^p\right\} \\ \le& \Pr\{\mathcal{N}_{s-1}^p \cap \neg\mathcal{N}_s^p\}.\end{aligned}$$

Then we prove that $\mathcal{N}_{s-1}^p \cap \neg\mathcal{N}_s^p \Rightarrow \neg\mathcal{N}_s^s$. In fact, if $\mathcal{N}_{s-1}^p$ happens, then we have $i^* \in S_{s-1}$, if $i^* \notin S_s$, then let $m_{s-1} = \arg\max_i \hat{\mu}_{i,s-1}$, we have

$$\begin{aligned}\mu_{i^*} - \hat{\mu}_{i^*,s-1} \ge& \mu_{m_{s-1}} - \hat{\mu}_{i^*,s-1} \\ \ge& \mu_{m_{s-1}} - \hat{\mu}_{m_{s-1},s-1} + 2\rho_s,\end{aligned}$$

which leads to $\neg\mathcal{N}_s^s$, since either $\mu_{i^*} - \hat{\mu}_{i^*,s-1} \ge \rho_s$ or $-\mu_{m_{s-1}} + \hat{\mu}_{m_{s-1},s-1} \ge \rho_s$ must happen. If $\mathcal{N}_{s-1}^p$ and $i^* \in S_s$ happens but $\exists i \in S_s, \mu_{i^*} - \mu_i > 4\rho_s$, then

$$\begin{aligned}&\mu_{i^*} - \hat{\mu}_{i^*,s-1} + \hat{\mu}_{i,s-1} - \mu_i \\ \ge& 4\rho_s - \hat{\mu}_{i^*,s-1} + \hat{\mu}_{i,s-1} \\ \ge& 4\rho_s - 2\rho_s \\ =& 2\rho_s,\end{aligned}$$

which also leads to $\neg\mathcal{N}_s^s$ by the same argument. So we have $\mathcal{N}_{s-1}^p \cap \neg\mathcal{N}_s^p \Rightarrow \neg\mathcal{N}_s^s$, then we have $\Pr\{\mathcal{N}_{s-1}^p \cap \neg\mathcal{N}_s^p\} \le P(\neg\mathcal{N}_s^s) \le \frac{2}{T^2}$ from the previous lemma,

$$\begin{aligned}\Pr\{\neg\mathcal{M}_t\} \le& \Pr\{\neg\mathcal{N}_1^p\} + \sum_{s=2}^{t} \Pr\{\mathcal{M}_{s-1} \cap \neg\mathcal{N}_s^p\} \\ \le& 0 + (t-1)\frac{2}{T^2} \\ \le& \frac{2}{T}.\end{aligned}$$

With the result of the previous lemma, we can prove Theorem 1. The proof is just a combination of Lemma 2 and direct calculation. We first partition the regret by an event $\mathcal{M}_T = \bigcap_{j=1}^{T} \mathcal{N}_j^p$, which is defined in Lemma 2, representing the event that for all $t \leq T$, the procedure is nice at round t. From Lemma 2, we know that the event will happen with high probability, and the regret in this case can be bounded easily. Then we just relax the regret in the case that $\mathcal{M}_T$ does not happen to the worst case and we will complete the proof. The proof of the theorem is straight forward, and we put the proof details in Appendix.

With Theorem 1, we can prove the regret for Algorithm 2. Direct computation will lead to Theorem 3. The detailed proof is shown in the Appendix.

3.3 Proof of Theorem 2

The proof of Theorem 2 is based on the following key observation. If arm j has mean value larger than that of arm $j+1$, i.e. $\mu_j \geq \mu_{j+1}$ and $\Delta_j \leq \Delta_{j+1}$, our algorithm will first play arm j and find that arm $j+1$ is bad and eliminate arm $j+1$. Then it will play arm j until arm j is eliminated by the algorithm. However, if we exchange arm j and arm $j+1$ such that in this case, $\mu_j < \mu_{j+1}$ and $\Delta_j < \Delta_{j+1}$, our algorithm will first play arm j for several times and find that arm j is bad and eliminate j, and then play arm $j+1$ until arm $j+1$ is eliminated. The number of total observations of arm j and arm $j+1$ is the same, but the regret of algorithm in the case of $\Delta_j < \Delta_{j+1}$ is worse then the case of $\Delta_j > \Delta_{j+1}$, because we spend more time playing the worse arm j in the first case. Therefore, the best sequence for our algorithms is $\Delta_1 \leq \Delta_2 \leq \cdots \leq \Delta_K$ with no regret, and the worst sequence is $\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_K$. Similarly, if i^* is the index of the best arm, when its index is fixed, for any sequence of arms before i^*, we can apply the above idea to do a bubble-sort on Δ_j's to change it into the worst sequence $\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_{i^*}$, and then use this worst sequence to bound the regret. In the following proof, we apply this bubble-sort idea to the proof of Lemma 3, which provides an upper bound to the optimal solution of a linear integer program. Then in the proof of Theorem 2, we show that the distribution-dependent regret is upper bounded by the optimal solution of the linear integer program.

Lemma 3. *Let $\{\Delta_{(i)}\}$ be a permutation of $\{\Delta_i \mid i \leq i^*\}$ such that $\Delta_{(1)} \geq \Delta_{(2)} \geq \cdots \geq \Delta_{(i^*)} = 0$, and let C be a constant. Then, let $(a_1, \ldots, a_{i^*}) \in \mathbb{N}^{i^*}$ denote the variables in the following optimization problem, the optimal value of the following optimization problem*

$$\max_{(a_1,\ldots,a_{i^*})\in\mathbb{N}^{i^*}} \sum_{j=1}^{i^*} a_j \Delta_j$$
$$s.t. \sum_{i=1}^{j} a_i \leq \frac{C}{\Delta_j^2} + 1, \forall j \in \{j' | a_{j'} > 0, j \neq i^*\},$$

is upper bounded by

$$\Delta_{(1)} + \frac{C}{\Delta_{(1)}} + C\sum_{i=2}^{i^*-1}\left(\frac{1}{\Delta_{(i)}^2} - \frac{1}{\Delta_{(i-1)}^2}\right)\Delta_{(i)}.$$

Proof. Let OPT denote the optimal value of the original optimization problem

$$\begin{aligned} \max_{(a_1,\ldots,a_{i^*})\in\mathbb{N}^{i^*}} & \sum_{j=1}^{i^*} a_j\Delta_j \\ \text{s.t.} & \sum_{i=1}^{j} a_i \le \frac{C}{\Delta_j^2} + 1, \forall j \in \{j' | a_{j'} > 0, j \ne i^*\}, \end{aligned} \tag{2}$$

and OPT' denote the optimal value of the modified optimization problem

$$\begin{aligned} \max_{(a_1,\ldots,a_{i^*})\in\mathbb{N}^{i^*}} & \sum_{j=1}^{i^*} a_j\Delta_{(j)} \\ \text{s.t.} & \sum_{i=1}^{j} a_i \le \frac{C}{\Delta_{(j)}^2} + 1, \forall j \in \{j' | a_{j'} > 0, j \ne i^*\}, \end{aligned} \tag{3}$$

where $\Delta_{(1)} \ge \Delta_{(2)} \ge \cdots \ge \Delta_{i^*}$ is a permutation of $\{\Delta_i\}_{i\le i^*}$. We first show that $OPT \le OPT'$. Suppose $\Delta_{j_0} < \Delta_{j_0+1}$. Let $\bar{\Delta}_{j_0} = \Delta_{j_0+1}, \bar{\Delta}_{j_0+1} = \Delta_{j_0}$, and for all $k \ne j_0, j_0+1$, $\bar{\Delta}_k = \Delta_k$,i.e. $\{\bar{\Delta}_j\}$ is obtained by exchanging 2 adjacent elements in $\{\Delta_j\}$. Let $\overline{OPT}$ denote the optimal value of the following optimization problem

$$\begin{aligned} \max_{(a_1,\ldots,a_{i^*})\in\mathbb{N}^{i^*}} & \sum_{j=1}^{i^*} a_j\bar{\Delta}_j \\ \text{s.t.} & \sum_{i=1}^{j} a_i \le \frac{C}{\bar{\Delta}_j^2} + 1, \forall j \in \{j' | a_{j'} > 0, j \ne i^*\}. \end{aligned} \tag{4}$$

We just have to show that $OPT \le \overline{OPT}$, then $OPT \le OPT'$ can be obtained by repeatly exchanging 2 adjacent elements. To prove $OPT \le \overline{OPT}$, we just have to show that every feasible solution in the original optimization problem (2) can be transformed into a feasible solution of the optimization problem (4), with the same objective value.

Let $x_1, \ldots, x_{i^*}$ be any feasible solution of the original optimization problem (2). Let $\bar{x}_j = x_j, \forall j \ne j_0, j_0+1$, and let $\bar{x}_{j_0} = x_{j_0+1}, \bar{x}_{j_0+1} = x_{j_0}$, and it is obvious that the objective value in the optimization problem (2) and (4) are the same, since we exchange the coefficient and the variable at j_0 and j_0+1 at the same time. Then we show that $\bar{x}_1, \ldots, \bar{x}_{i^*}$ is also a feasible solution in optimization problem (4).

First for all $j \neq j_0, j_0+1$, we have $\sum_{i=1}^{j} \bar{x}_i \leq \frac{C}{\bar{\Delta}_j^2} + 1$, since it is equivalent to $\sum_{i=1}^{j} x_i \leq \frac{C}{\Delta_j^2} + 1$ and $\bar{x}_j > 0$ is equivalent to $x_j > 0$.

Then we consider the variable $\bar{x}_{j_0} = x_{j_0+1}$. If $x_{j_0+1} > 0$, we have

$$\sum_{i=1}^{j_0} \bar{x}_i \leq \sum_{i=1}^{j_0+1} \bar{x}_i = \sum_{i=1}^{j_0+1} x_i \leq \frac{C}{\Delta_{j_0+1}^2} + 1 = \frac{C}{\bar{\Delta}_{j_0}^2} + 1.$$

If $x_{j_0+1} = 0$, then $\bar{x}_{j_0} = 0$ and we do not have a constraint for $j = j_0$ in problem (4).

Next we consider the variable $\bar{x}_{j_0+1} = x_{j_0}$. If $x_{j_0+1} > 0$, using $\bar{\Delta}_{j_0+1} < \bar{\Delta}_{j_0}$ we have

$$\sum_{i=1}^{j_0+1} \bar{x}_i = \sum_{i=1}^{j_0+1} x_i \leq \frac{C}{\Delta_{j_0+1}^2} + 1 = \frac{C}{\bar{\Delta}_{j_0}^2} + 1 \leq \frac{C}{\bar{\Delta}_{j_0+1}^2} + 1.$$

If $x_{j_0+1} = 0$ and $x_{j_0} > 0$, we have

$$\sum_{i=1}^{j_0+1} \bar{x}_i = \sum_{i=1}^{j_0+1} x_i = \sum_{i=1}^{j_0} x_i \leq \frac{C}{\Delta_{j_0}^2} + 1 = \frac{C}{\bar{\Delta}_{j_0+1}^2} + 1.$$

If $x_{j_0} = 0$, then $\bar{x}_{j_0+1} = 0$ and we do not need a constraint for $j = j_0 + 1$ in problem (4).

Therefore, after discussing all cases, we know that $(\bar{x}_1, \ldots, \bar{x}_K)$ is a feasible solution of the optimization problem (4). Then with our previous argument, the optimal value OPT' of optimization problem (3) is at least OPT, i.e. $OPT \leq OPT'$.

Then suppose $\{x_{r_i}\}$ is a feasible solution of the modified optimization problem (3), we have

$$\begin{aligned}
\sum_{i=1}^{i^*} x_{r_i} \Delta_{(i)} =& x_{r_1} \Delta_{(1)} + \sum_{i=2}^{i^*} \left(\sum_{j=1}^{i} x_{r_j} - \sum_{j=1}^{i-1} x_{r_j} \right) \Delta_{(i)} \\
=& \sum_{i=1}^{i^*-1} \left(\Delta_{(i)} - \Delta_{(i+1)} \right) \sum_{j=1}^{i} x_{r_j} + \Delta_{(i^*)} \sum_{j=1}^{i^*} x_{r_j} \\
\leq& \sum_{i=1}^{i^*-1} \left(\Delta_{(i)} - \Delta_{(i+1)} \right) \left(\frac{C}{\Delta_{(i)}^2} + 1 \right) \\
=& \Delta_{(1)} + \frac{C}{\Delta_{(1)}} + C \sum_{i=2}^{i^*-1} \left(\frac{1}{\Delta_{(i)}^2} - \frac{1}{\Delta_{(i-1)}^2} \right) \Delta_{(i)},
\end{aligned}$$

where we use the fact that $\Delta_{(i^*)} = 0$. So OPT' is also upper bounded, which complete the proof directly. □

With the conclusion of the lemma, we can prove Theorem 2. The general idea to prove Theorem 2 is the same as proving Theorem 1. We first partition the regret by the event $\mathcal{M}_T$, which is defined in Definition 4. With Lemma 2, $\mathcal{M}_T$ will happen with high probability, and we can just consider the regret when $\mathcal{M}_T$ happens. Then we bound the regret when $\mathcal{M}_T$ happens from the help of Lemma 3.

Proof (Proof of Theorem 2). Similar to the proof of Theorem 1, we have

$$\mathbb{E}[\sum_{t=1}^{T}(\mu_{i^*}-\mu_{I_t})] \leq \mathbb{E}[\sum_{t=1}^{T}(\mu_{i^*}-\mu_{I_t})|\mathcal{M}_T] + \mathbb{E}[\sum_{t=1}^{T}(\mu_{i^*}-\mu_{I_t})|\neg\mathcal{M}_T]\cdot\Pr\{\neg\mathcal{M}_T\},$$

and

$$\mathbb{E}[\sum_{t=1}^{T}(\mu_{i^*}-\mu_{I_t})|\neg\mathcal{M}_T] \leq T, \Pr\{\neg\mathcal{M}_T\} \leq \frac{2}{T}.$$

Suppose that the arms $1, 2, \ldots, K$ are played for $a_1, a_2, \ldots, a_K$ times after T rounds and $\mathcal{M}_T$ happens, then the regret is $\sum_{i=1}^{K} a_i \Delta_i$. Then we show that when $\mathcal{M}_T$ happens,

$$\sum_{i=1}^{K} a_i\Delta_i \leq \Delta_{(1)} + \frac{C}{\Delta_{(1)}} + C\sum_{i=2}^{i^*-1}\left(\frac{1}{\Delta_{(i)}^2} - \frac{1}{\Delta_{(i-1)}^2}\right)\Delta_{(i)}.$$

First, we only have to consider the arm j with $j < i^*$, since if $\mathcal{M}_T$ happens, our elimination based algorithm (see Algorithm 1) will never choose arm $j > i$ to play, and for arm i^* there is no regret contribution. For arm $j < i^*$, if $a_j \neq 0$, then at the last time the algorithm plays arm j, arm j has been observed for $\sum_{i=1}^{j} a_j - 1$ times, since we only delete the arms in set S so I_t must be non-decreasing. Then as $\mathcal{M}_T$ happens, we have

$$\Delta_j \leq 4\sqrt{\frac{\ln(KT^2)}{2(\sum_{i=1}^{j} a_i - 1)}},$$

which will lead to

$$\sum_{i=1}^{j} a_i \leq \frac{C}{\Delta_j^2} + 1,$$

where $C = 8\ln(KT^2)$ as defined in Theorem 2. Then we can conclude that when $\mathcal{M}_T$ happens, the regret is bounded by

$$\max_{(a_1,\ldots,a_{i^*})\in\mathbb{N}^{i^*}} \sum_{j=1}^{i^*-1} a_j\Delta_j$$
$$\text{s.t.} \sum_{i=1}^{j} a_i \leq \frac{C}{\Delta_j^2} + 1, \forall j \in \{j' | a_{j'} > 0, j \neq i^*\}.$$

Then from Lemma 3, we know that the optimal value of the above optimization problem is upper bounded, so we have

$$\mathbb{E}[\sum_{t=1}^{T}(\mu_{i^*} - \mu_{I_t})|\mathcal{M}_T] \leq \Delta_{(1)} + \frac{C}{\Delta_{(1)}} + C\sum_{i=2}^{i^*-1}\left(\frac{1}{\Delta_{(i)}^2} - \frac{1}{\Delta_{(i-1)}^2}\right)\Delta_{(i)}$$

Then combine with the previous result, we can finish the proof. □

Then we have a corollary from this distribution dependent bound.

Corollary 1. *Let $\{\Delta_{(i)}\}$ be a permutation of $\{\Delta_i\}$ such that $\Delta_{(1)} \geq \Delta_{(2)} \geq \cdots \geq \Delta_{(i^*)} = 0$, and $C = 8\ln(KT^2)$ as defined in Theorem 2, then the regret is bounded by $\left(\frac{C}{\Delta_{(i^*-1)}^2} + 1\right)\Delta_{(1)} + 2$.*

4 Lower Bound

The lower bound for multi-armed bandit problems has been extensively studied. However, we notice that there is no regret lower bound for the full-information multi-armed bandit under the stochastic case. In this section, we show that the regret is lower bounded by $\Omega(\sqrt{T\log K})$ in this case, which also implies a regret lower bound of $\Omega(\sqrt{T\log K})$ for the one-sided bandit case. Comparing with the regret upper bound of $O(\sqrt{T(\log K + \log T)})$ of Theorem 1, we can see that our elimination algorithm gives almost a tight regret bound.

In this section, we fix a bandit algorithm. Let I_t denote the choice of the algorithm in round t. Let K denote the total number of arms. For each $j \in [K]$, let $\mathcal{I}_j$ denote the problem instance that $\mu_k = \frac{1}{2}$, for all $k \neq j$, $\mu_j = \frac{1+\varepsilon}{2}$ for some small $\varepsilon > 0$, and each arm is a Bernoulli random variable independent from other arms.

The proof follows from the original proof of lower bound for bandit feedback MAB problem [5], but we need more careful calculation. The original proof for the bandit feedback regret lower bound is $\sqrt{TK}$, and if we directly apply it to the full information feedback case, we would get $\sqrt{T}$ lower bound. With more careful analysis, we could raise this lower bound to $\sqrt{T\log K}$. Following the original analysis, we connect the full information MAB problem with the *bandit-with-prediction* problem, in which the algorithm is given the rewards of all arms in the first T rounds, and it needs to decide which is the best arm. We use y_T to denote the output of an algorithm of the bandit-with-prediction problem in this section. Naturally, we can select the arm with the largest cumulative rewards in the first T rounds as y_T, and this is called Follow-the-Leader strategy. Then we use the reverse Chernoff Bound (Lemma 4) to show the regret lower bound for the Follow-the-Leader strategy, and then we show that Follow-the-Leader strategy has the optimal regret among all the algorithm (up to constants). Finally, we reduce the full information MAB problem to the bandit-with-prediction problem to show its lower bound.

Lemma 4. (Tightness of Chernoff Bound) *Suppose $X_1, X_2, \ldots, X_n$ are i.i.d Bernoulli random variable with $Pr[X_1 = 1] = \frac{1}{2}$, then there exists absolute constants c', d, p such that for all $0 < \varepsilon < d$ such that $\varepsilon^2 \cdot n > p$,*

$$Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > \frac{1}{2} + \varepsilon\right\} > e^{-c'n\varepsilon^2}.$$

The above lemma is a well-known result. For convenience, we put the proof of this lemma in the appendix. Please see Appendix for more details. The following lemma shows that the Follow-the-Leader strategy still could make mistakes on the bandit-with-prediction task.

Lemma 5. *Suppose $\frac{c\ln K}{2\varepsilon^2} \leq T \leq \frac{c\ln K}{\varepsilon^2}$, for a small enough absolute constant c (which is not the constant in the previous lemma) and $0 \leq \varepsilon < d$ (where d is the absolute constant in the previous lemma). Consider the algorithm Follow-the-Leader for the bandits-with-prediction problem. Then for large enough K,*

$$\sum_{j=1}^{K} Pr\{y_T = j|\mathcal{I}_j\} \leq \frac{K}{4}.$$

The next lemma shows that no other algorithms can do much better than the Follow-the-Leader strategy, for the bandit-with-prediction problem.

Lemma 6. *Suppose $\frac{c\ln K}{2\varepsilon^2} \leq T \leq \frac{c\ln K}{\varepsilon^2}$, for a small enough absolute constant c, a large enough K and $0 \leq \varepsilon < d$ where d is the constant in previous lemma. Then for any (deterministic or randomized) algorithm for the bandit-with-prediction problem, there exists at least $\lceil K/3 \rceil$ arms j such that*

$$Pr\{y_T = j|\mathcal{I}_j\} \leq \frac{3}{4}.$$

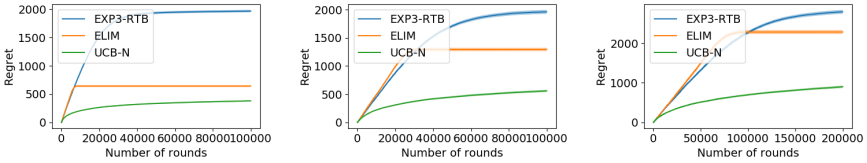
We can now prove the regret lower bound of the full information bandit problem by utilizing the above result for the bandit-with-prediction problem.

Theorem 4. (Regret lower bound for full information stochastic bandits) *Fix time horizon T and the number of arms K such that $\sqrt{c\ln K/T} < d$, where c, d are the constants in Lemma 6. When K is big enough, then for any bandit algorithm, there exists a problem instance such that $\mathbb{E}[R(T)] \geq \Omega(\sqrt{T\log K})$.*

Please see Appendix for the missing proofs.

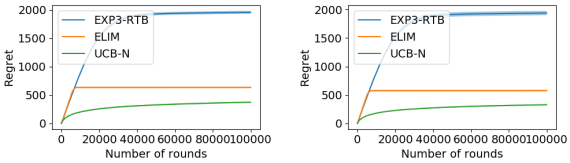
5 Numerical Experiments

In this section, we show numerical experiments on our elimination based algorithm ELIM together with two other algorithms: (a) EXP3-RTB algorithm



(a) $\Delta = 0.1$ (b) $\Delta = 0.05$ (c) $\Delta = 0.03$

Fig. 1. Uniform-mean suboptimal arms with $K = 20$, $Best = 17$, and varying Δ.



(a) $Best = 10$ (b) $Best = 3$

Fig. 2. Uniform-mean suboptimal arms with $K = 20$, $\Delta = 0.1$, and varying $Best$.

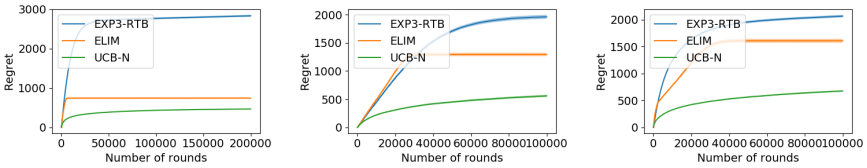
introduced in [7], which solves one-sided full information bandit in the adversarial case, and (b) UCB-N algorithm introduced in [6] to solve stochastic multi-armed bandit with side information, and it is essentially UCB but updates any arm when it has an observation, not just the arm played in the round.

First, we do experiments when all the suboptimal arms have the same mean of 0.6 with a gap Δ towards the best arm, similar to our lower bound analysis setting. We will show results with different Δ setting (Fig.1) and different best arm position (Fig.2). For convenience, we let the reward of each arm follows a Bernoulli distribution. Next, we do experiments when the suboptimal arms have means drawn uniformly at random from $(0.2, 0.6)$ except for the mean of the best arm, which is set to $0.6 + \Lambda$ for a parameter Λ. We vary the value of Λ (Fig.3) and the position of the best arms (Fig.4).

We use T to denote the total time horizon we choose in the experiments. In most of the experiments, we choose $T = 100000$, but we will choose $T = 200000$ to better distinguish the performance between different algorithms in some cases. We use K to denote the number of arms in our experiments, and we choose $K = 20$ in all of the experiments. We use $Best$ to denote the position of the best arm, which is set to $3, 10, 17$ in different experiments. For each experiment, we run 100 times and draw the 99% confidence interval surrounding the curve (all are very narrow regions surrounding the curve).

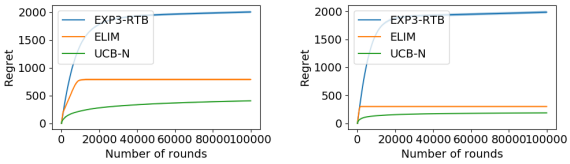
From the above experiments, we can find that

1. In both experiments, when the gap between the mean of the best arm and the mean of others is larger, our algorithm performs much better than the existing EXP3-RTB algorithm.



(a) $\Lambda = 0.1$ (b) $\Lambda = 0.05$ (c) $\Lambda = 0.03$

Fig. 3. Random-mean suboptimal arms with $K = 20$, $Best = 17$, and varying Δ.



(a) $Best = 10$ (b) $Best = 3$

Fig. 4. Uniform-mean suboptimal arms with $K = 20$, $\Delta = 0.1$, and varying $Best$.

2. In the first experiments, when we change the position of the best arm, the regret line does not change so much. In the second experiments where we add more randomness, if the position of the best arm has small index, then our algorithm will perform better. However, the existing EXP3-RTB algorithm does not have this property.
3. UCB-N consistently outperforms both our algorithm ELIM and the EXP3-RTB algorithm.

Therefore, we can conclude in the stochastic setting, our elimination-based algorithm performs much better than the EXP3-RTB algorithm designed for the same problem but on the adversarial setting, and UCB-N has the best empirical performance. The issue with UCB-N is that we cannot derive a tight theoretical regret bound that also beats or even match ELIM. If we simply use UCB regret bound for UCB-N, it would be too loose and it would be inferior to our elimination based algorithm, as discussed after Theorem 2. The result in [6] on UCB-N cannot be applied here either because it requires mutually observable cliques in the observation graph but for the one-sided full-information case, the only cliques are the trivial singletons, which makes their regret bound reduced to the UCB regret bound. Therefore, our algorithm ELIM is the one that achieves the best theoretical regret bound, significantly outperform the EXP3-RTB algorithm for the adversarial case, while UCB-N has the best empirical performance with an unknown tight theoretical guarantee.

6 Conclusion and Further Work

In this paper, we study the stochastic one-sided full-information bandit and propose an elimination-based algorithm to solve the problem. We provide the upper bounds of the algorithm, and show that it almost matches the lower bound of the problem. Our experiment demonstrates that it performs better than the algorithm designed for the adversarial setting. To the best of our knowledge, our algorithm achieves the best regret bound so far.

One open problem is definitely on the analysis of UCB-N. As we have discussed, its naive regret bound such as the UCB regret bound would be much worse than our elimination algorithm, but its empirical performance shows better results. We are trying to provide a tighter analysis on UCB-N, but it evades several attempts we have made so far, and thus we left it as a future research question. Another direction is to design other algorithms that better utilizes the one-sided full-information feedback structure and achieves both good theoretical and empirical results. Other specific feedback structures corresponding to practical applications are also worth further investigation.

7 Acknowledgement

Wei Chen is partially supported by the National Natural Science Foundation of China (Grant No. 61433014).

References

1. Alon, N., Cesa-Bianchi, N., Gentile, C., Mansour, Y.: From bandits to experts: A tale of domination and independence. In: Advances in Neural Information Processing Systems. pp. 1610–1618 (2013)
2. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine Learning **47**(2-3), 235–256 (2002)
3. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multi-armed bandit problem. SIAM J. Comput. **32**(1), 48–77 (2002)
4. Berry, D.A., Fristedt, B.: Bandit problems: Sequential Allocation of Experiments. Chapman and Hall (1985)
5. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends® in Machine Learning **5**(1), 1–122 (2012)
6. Caron, S., Kveton, B., Lelarge, M., Bhagat, S.: Leveraging side observations in stochastic bandits. In: UAI. pp. 142–151 (2012)
7. Cesa-Bianchi, N., Gaillard, P., Gentile, C., Gerchinovitz, S.: Algorithmic chaining and the role of partial feedback in online nonparametric learning. In: Proceedings of the 30th Conference on Learning Theory. pp. 465–481 (2017)
8. Mannor, S., Shamir, O.: From bandits to experts: On the value of side-observations. In: Advances in Neural Information Processing Systems. pp. 684–692 (2011)
9. Robbins, H.: Some aspects of the sequential design of experiments. Bulletin American Mathematical Society **55**, 527–535 (1952)
10. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**(3/4), 285–294 (1933)