Cost-Aware Cascading Bandits

Chao Gan, Ruida Zhou, Jing Yang, *Member, IEEE*, and Cong Shen, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a cost-aware cascading bandits model, a new variant of multi-armed bandits with cascading feedback, by considering the random cost of pulling arms. In each step, the learning agent chooses an *ordered* list of items and examines them sequentially, until certain stopping condition is satisfied. Our objective is then to maximize the expected *net reward* in each step, i.e., the reward obtained in each step minus the total cost incurred in examining the items, by deciding the ordered list of items, as well as when to stop examination. We first consider the setting where the instantaneous cost of pulling an arm is unknown to the learner until it has been pulled. We study both the offline and online settings, depending on whether the state and cost statistics of the items are known beforehand. For the offline setting, we show that the Unit Cost Ranking with Threshold 1 (UCR-T1) policy is optimal. For the online setting, we propose a Cost-aware Cascading Upper Confidence Bound (CC-UCB) algorithm, and show that the cumulative regret scales in $O(\log T)$. We also provide a lower bound for all α-consistent policies, which scales in $\Omega(\log T)$ and matches our upper bound. We then investigate the setting where the instantaneous cost of pulling each arm is available to the learner for its decision-making, and show that a slight modification of the CC-UCB algorithm, termed as CC-UCB2, is order-optimal. The performances of the algorithms are evaluated with both synthetic and real-world data.**

***Index Terms*—Cost-aware, cascading bandits, upper confidence bound.**

I. Introduction

In this paper, we introduce a new cost-aware cascading bandits (CCB) model. We consider a set of K items (arms) denoted as $[K] = \{1, 2, \ldots, K\}$. Each item $i \in [K]$ has two possible states 0 and 1, which evolve according to an independent and identically distributed (i.i.d.) Bernoulli random variable. The learning agent chooses an *ordered* list of items in each step and examines them sequentially until certain stopping condition is met. The reward that the learning agent receives in a step equals one if one of the examined items in that step has state 1; Otherwise, it equals zero. We associate a random cost for examining each item. The overall reward function, termed as *net reward*, is the reward obtained in each step minus the total cost incurred in examining the items before the learner stops.

The CCB model is a natural but technically non-trivial extension of cascading bandits [2], and is a more suitable model in many fields, including the following examples.

1) *Opportunistic spectrum access:* In cognitive radio systems, a user is able to probe multiple channels sequentially at the beginning of a transmission session before it decides to use at most one of the channels for data transmission. Assuming the channel states evolve between *busy* and *idle* in time, the user gets a reward if there exists one *idle* channel for transmission. The cost then corresponds to the energy and/or delay incurred in probing each channel.
2) *Dynamic treatment allocation:* In clinic trials, a doctor must assign one of several treatments to a patient in order to cure a disease. The doctor accrues information about the outcome of previous treatments before making the next assignment. Whether a treatment cures the disease can be modeled as a Bernoulli random variable, and the doctor gets a reward if the patient is cured. The doctor may not only be interested in the expected effect of the treatment but also its riskiness, which can be interpreted as the cost of the treatment.

In both examples, the net reward in each step is determined not only by the subset of items included in the list, but also by the *order* that they are examined. Intuitively, if the costs in examining the items are homogeneous, we would prefer to have the channel with higher probability to be idle, or the more effective treatment, ranked higher in the list. Then, the learner would find an available channel or cure the patient after a few attempts and then stop examination, thus saving the cost without decreasing the reward. However, for more general cases where the costs are heterogeneous or even random, the optimal solution is not immediately clear. This has motivated us to consider a general cost model where the costs of pulling arms are heterogeneous and random, and investigate the corresponding solutions.

A. Main Contributions

Our main contributions are four-fold:

1) We propose a novel CCB model, which has implications in many practical scenarios, such as opportunistic spectrum access and dynamic treatment allocation. The CCB model is fundamentally different from its cost-oblivious counterparts, and admits a unique structure in the corresponding learning strategy.

Manuscript received July 18, 2019; revised March 31, 2020 and May 19, 2020; accepted May 29, 2020. Date of publication June 10, 2020; date of current version June 26, 2020. The associate editor coordinating the reviewof this manuscript and approving it for publication was Dr. Caroline Chaux. This work was supported in part by the US National Science Foundation under Grant ECCS-1650299. It was presented in part in the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018 [1]. *(Chao Gan and Ruida Zhou contributed equally to this work.) (Corresponding author: Chao Gan.)*

Chao Gan and Jing Yang are with the School of Electrical Engineering and Computer Science, The Pennsylvania State University, University Park, PA 16801 USA (e-mail: cug203@psu.edu; yangjing@psu.edu).

Ruida Zhou is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: ruida@tamu.edu).

Cong Shen is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: cshen@ieee.org).

Digital Object Identifier 10.1109/TSP.2020.3001388

2) Second, we explicitly identify the special structure of the optimal offline policy, which serves as the baseline for the online algorithms we develop. The optimal offline policy, called Unit Cost Ranking with Threshold 1 (UCR-T1), is to form an *ordered* list of arms by including the arms whose reward-cost ratio exceeds threshold 1, and pull them sequentially until a state 1 is observed or all arms are pulled.
3) Third, we propose a cost-aware cascading Upper Confidence Bound (CC-UCB) algorithm for the scenario when prior arm statistics are unavailable, and show that it is order-optimal by establishing order-matching upper and lower bounds on the regret. Our analysis indicates that the UCB based algorithm performs well for ranking the arms, i.e., the cumulative regret of ranking the desired arms in a wrong order is bounded.
4) Finally, we extend the analysis to the setting where the instantaneous costs of pulling individual arms are available to the learner for its decision-making. We modify the CC-UCB algorithm to track the time-varying optimal offline policy, and show that the cumulative regret has different dependency on the arms included in certain optimal lists and those that are never included in any optimal list. When all arms are included in certain optimal lists, the regret can be bounded by a constant. Lower bound also matches with the corresponding upper bound.

B. Relation to the State of the Art

Bandits with cost and budget considerations: There have been some attempts to take the cost of pulling arms and budget constraint into the multi-armed bandit (MAB) framework recently. They can be summarized in two types. In the first type [3]–[7], pulling each arm in the exploration phase has a unit cost and the goal is to find the best arm given the budget constraint on the total number of exploration arms. This type of problems is also referred to as “best-arm identification” or “pure exploration”. In the second type, pulling an arm is always associated with a cost and constrained by a budget, no matter in the exploration phase or the exploitation phase, and the objective usually is to design an arm pulling algorithm in order to maximize the total reward with given cost or budget constraint. References [8]–[10] consider the problem when the cost of pulling each arm is *fixed* and becomes known after the arm is used once. A sample-path cost constraint with *known* bandit dependent cost is considered in [11]. References [12]–[15] study the budgeted bandit problems with *random* arm pulling costs. Reference [16] considers the knapsack problem where there can be more than one budget constraints and shows how to construct polices with sub-linear regret.

In the proposed CCB model, the *net reward* function is related to the cost of pulling arms and the learning agent faces a “soft constraint” on the cost instead of a fixed budget constraint. If the learner only pulls one arm in each step, the cost of pulling an arm can be absorbed into the reward of that arm (i.e., net reward). Our model then reduces to a conventional MAB model for this case. In this paper, however, we are interested in the scenario where the learner is allowed to sequentially pull a number of arms in each step, and the reward obtained in each step cannot be decomposed into the summation of the rewards from the pulled arms. Thus, the cost cannot be simply absorbed into the reward of individual arms. The intricate relationship between the cost and the reward, and its implications on the optimal policy require more sophisticated analysis, and that is the focus of this paper.

The cost of arm pulling is related to certain risk and safety concerns in the bandit literature, where various risk measures and risk-averse bandits have been considered [17]–[19]. The main difference between the setups of risk-averse MAB and the cost-aware MAB is, the risk is often related to the distribution of the rewards, while the cost is usually assumed to be independent of the rewards.

Bandits with multiple pulls each time: MAB with more than one arm to pull in each step has been studied in multiple-play MAB (MP-MAB) models in [20]–[22], cascading bandits (CB) models in [2], [23], [24], and ranked bandits (RB) models in [25]–[27]. Under MP-MAB, the learner is allowed to pull L out of K arms, and the reward equals to the summation of the rewards from individual arms. Although the proposed CCB model also allows the user to pull multiple arms in each step, the reward is not accumulative, thus leading to different solutions.

The CCB model proposed in this paper is closely related to the CB model investigated in [2]. Specifically, [2] considers the scenario where at each step, L out of K items are listed by a learning agent and presented to a user. The user examines the ordered list from the first to the last, until he/she finds the first attractive item and clicks it. The system receives a reward if the user finds at least one of the items to be attractive. Our model has the same reward model as that in the cascading bandits setting. However, there are also important distinctions between them. In the CB model in [2], the total number of items to be examined each step is fixed, and the cost of pulling individual arms is not considered. As a result, the same subset of items on a list will give the same expected reward, irrespective of their order on the list. However, for the proposed CCB model, the ranking of the items on a list does affect the expected net reward. Therefore, the structure of the optimal offline policy and the online algorithm we develop are fundamentally different from those in [2].

The proposed CCB model is also related to RB in the sense that the order of the arms to pull in each step matters. A crucial feature of RB is that the click probability for a given item may depend on the item and its position on the list, as well as the items shown above. However, in our case, we assume the state of an arm evolves in an i.i.d. fashion from step to step.

C. Paper Outline

This paper is organized as follows. Section II describes the system model. Section III and Section IV describe the known and unknown instantaneous cost cases, respectively. Section V discusses a fail-safe extension of the CCB model, and Section VI evaluates the proposed algorithms through simulations. Concluding remarks are provided in Section VII. Important proofs are deferred to the Appendix.

II. System Model

Consider a K-armed stochastic bandit system where the state of each arm evolves independently from step to step. Let $X_{i,t}$ be the state of arm $i \in [K]$ in step t. Then, $X_{i,t} \in \{0,1\}$ evolves according to an i.i.d. Bernoulli distribution with $\mathbb{E}[X_{i,t}] = \theta_i$ in all steps t. Denote $Y_{i,t}$ as the cost of pulling arm i in step t, where $Y_{i,t}$ is a bounded and non-negative i.i.d. random variable with $\mathbb{E}[Y_{i,t}] = c_i$. We assume the expected value of $X_{i,t}$ (i.e., θ_i) is unknown beforehand, and the specific value of $X_{i,t}$ is revealed to the learner once it is pulled. For the cost $Y_{i,t}$, we consider two different settings. For the first setting, similar to $X_{i,t}$, we assume the expected value of $Y_{i,t}$ (i.e., c_i) is unknown a priori, and $Y_{i,t}$ is observed once arm i is pulled. For the second setting, we assume the instantaneous costs $\{Y_{i,t}\}_{i\in[K]}$ are available to the learner at the beginning of step t.

In step t, the learning agent chooses an ordered list of arms from $[K]$ and pulls the arms sequentially. Denote the ordered list as $I_t := \{I_t(1), I_t(2), \ldots, I_t(|I_t|)\}$, where $I_t(i)$ is the ith arm to be pulled, and $|I_t|$ is the cardinality of I_t. Denote $\tilde{I}_t$ as the list of arms that have been actually pulled in step t. We have $\tilde{I}_t \subseteq I_t$.

The *reward* that the learning agent receives in step t depends on both $\tilde{I}_t$ and $\{X_{i,t}\}_{i\in\tilde{I}_t}$. Specifically, it can be expressed as $1 - \prod_{i=1}^{|\tilde{I}_t|}(1 - X_{\tilde{I}_t(i),t})$, i.e., the learning agent gets reward one if one of the arms that have been examined in step t has state 1; Otherwise, it equals zero. The *cost* that is incurred at step t also depends on $\tilde{I}_t$, and it can be expressed as $\sum_{i=1}^{|\tilde{I}_t|} Y_{\tilde{I}_t(i),t}$.

With a given ordered list I_t, $\tilde{I}_t$ is random and its realization depends on the observed $X_{i,t}, Y_{i,t}$ and the stopping condition in general. Denote the *net reward* received by the learning agent at step t as $r_t := 1 - \prod_{i=1}^{|\tilde{I}_t|}(1 - X_{\tilde{I}_t(i),t}) - \sum_{i=1}^{|\tilde{I}_t|} Y_{\tilde{I}_t(i),t}$.

Define the *per-step regret* $\mathbf{reg}_t := r_t^* - r_t$, where r_t^* is the *net reward* that would be obtained at step t if the corresponding statistics were known beforehand and the optimal I_t and stopping condition were adopted. When $Y_{i,t}$ is observed only after it is pulled and the statistics $\{c_i\}_i$ are unknown before hand, r_t^* depends on $\{\theta_i\}_i$ and $\{c_i\}_i$.

Without prior knowledge of $\{\theta_i\}_i$ or $\{c_i\}_i$, our goal is to design an online algorithm to decide I_t based on previous observations $\cup_{\tau=1}^{t-1}\{X_{i,\tau}, Y_{i,\tau}\}_{i\in\tilde{I}_\tau}$, and $\tilde{I}_t$ based on observed states and costs in step t, so as to minimize the *cumulative regret* $R(T) := \mathbb{E}[\sum_{t=1}^{T}\mathbf{reg}_t]$.

On the other hand, if $\{Y_{i,t}\}_{i\in[K]}$ are available to the learner at the beginning of step t, r_t^* depends on $\{\theta_i\}_i$ and $\{Y_{i,t}\}_{i\in[K]}$. Without prior knowledge of $\{\theta_i\}_i$, our goal is to minimize the cumulative regret by deciding I_t based on $\cup_{\tau=1}^{t-1}\{X_{i,\tau}\}_{i\in\tilde{I}_\tau}$ and the instantaneous costs $\{Y_{i,t}\}_{i\in[K]}$, and $\tilde{I}_t$ based on observed $\{X_{i,t}\}$.

In the following, we will consider those two different but closely related settings separately. For both settings, we will first identify the structure of the corresponding optimal offline policy, and then develop an online algorithm to learn the statistics and track the optimal offline policy progressively.

III. Unknown Immediate Costs

In this section, we consider the setting where the cost of pulling each arm $Y_{i,t}$ is unknown until it is pulled. Assuming the statistics of the rewards and costs are unknown a priori, the learner needs to learn them through the observed arm states $\{X_{i,t}\}$ and costs $\{Y_{i,t}\}$.

A. Optimal Offline Policy

We first study the optimal offline policy for the non-trivial case where $c_i > 0$. The offline setting corresponds to the case that the arm statistics $\{\theta_i\}_{i=1}^K$ and $\{c_i\}_{i=1}^K$ are known to the learning agent as prior knowledge. However, the instantaneous realization of rewards and costs associated with arms are unknown to the learning agent until the arms are actually pulled. Under the assumption that the distributions of arm states and costs are i.i.d. across steps, the optimal offline policy should remain the same for different steps. Thus, in this section, we drop the step index and focus on the policy at an individual step.

According to the definition of the cascading feedback, for any ordered list, the reward in each step will not grow after the learner observes an arm with state 1. Therefore, *to maximize the net reward, the learner should stop examining the rest of the list when a state 1 is observed, in order to save the cost of examination*. Let the ordered list under the optimal offline policy be I^*, then

$$I^* = \arg\max_I \sum_{i=1}^{|I|}(\theta_{I(i)} - c_{I(i)}) \prod_{j=1}^{i-1}(1 - \theta_{I(j)}), \quad (1)$$

$$\mathbb{E}[r^*] = \sum_{i=1}^{|I^*|}(\theta_{I(i)} - c_{I(i)}) \prod_{j=1}^{i-1}(1 - \theta_{I(j)}). \quad (2)$$

We note that the expected net reward structure is more complex than the standard multi-armed bandits problem or the standard cascading model, and the optimal offline policy is not straightforward. By observing (2), we note that there are both a subtraction part $\theta_{I(i)} - c_{I(i)}$ and a product part $\prod_{j=1}^{i-1}(1 - \theta_{I(j)})$ inside each summation term. On one hand we should choose large θ_i to increase the value of $\theta_i - c_i$, but on the other hand large θ_i contributes to smaller $1 - \theta_i$. In the extreme case where no cost is assigned to arms, the optimal policy is to pull all arms in *any* order, and the problem becomes trivial.

For simplicity of the analysis, we make the following assumptions:

Assumptions 1: 1) $\theta_i \neq c_i$, for all $i \in [K]$. 2) There exists a constant $\epsilon > 0$, such that $c_i > \epsilon$ for all $i \in [K]$.

Under Assumptions 1, we present the optimal offline policy in Theorem 1, which is called Unit Cost Ranking with Threshold 1 (UCR-T1), as it ranks the expected reward normalized by the average cost and compares against threshold one. The violation of Assumption 1.1 would make the optimal offline policy non-unique.

Theorem 1: Arrange the arm indices such that

$$\frac{\theta_{1^*}}{c_{1^*}} \geq \frac{\theta_{2^*}}{c_{2^*}} \geq \cdots \geq \frac{\theta_{L^*}}{c_{L^*}} > 1 > \frac{\theta_{(L+1)^*}}{c_{(L+1)^*}} \geq \cdots \geq \frac{\theta_{K^*}}{c_{K^*}}.$$

Then, $I^* = \{1^*, 2^*, \ldots, L^*\}$, and the corresponding optimal expected per-step net reward is $\mathbb{E}[r^*] = \sum_{i=1}^{L}(\theta_{i^*} - c_{i^*})\prod_{j=1}^{i-1}(1-\theta_{j^*})$.

The proof of Theorem 1 is provided in Appendix A. Theorem 1 indicates that ranking the arms in a descending order of $\frac{\theta_i}{c_i}$ and only including those with $\frac{\theta_i}{c_i} > 1$ in I^* achieves a balanced tradeoff between maximizing the expected net reward from the current arm (i.e., $\theta_i - c_i$) and maximizing the cumulative net reward from the remaining arms in I^* if the current arm has state 0. This is an important observation which will also be useful in the online policy design.

B. Online Algorithm and Upper Bound

With the optimal offline policy explicitly described in Theorem 1, in this section, we develop an online algorithm to maximize the cumulative expected net rewards without a priori knowledge of $\{\theta_i\}_{i=1}^K$ and $\{c_i\}_{i=1}^K$.

Unlike the previous work on MAB, the net reward structure in our setting is rather complex. One difficulty is that the learner has to rank θ_i/c_i and compare it with threshold 1 for exploitation. A method to deal with this difficulty is using a UCB-type algorithm following the Optimism in Face of Uncertainty (OFU) principle [28]. More specifically, we use a UCB-type indexing policy to rank the arms. Though the upper confidence bound is a biased estimation of θ_i/c_i, for any arm in I^*, it will converge to the true value asymptotically.

The cost-aware cascading UCB (CC-UCB) algorithm is described in Algorithm 1. The costs are assumed to be random but the learning agent has no knowledge of their distributions. We use $N_{i,t}$ to track the number of steps that arm i has been pulled right before step t, and $\hat{\theta}_{i,t}$, $\hat{c}_{i,t}$ to denote the sample average of θ_i and c_i at step t, respectively. The UCB padding term on the state and cost of arm i at step t is $u_{i,t} := \sqrt{\frac{\alpha \log t}{N_{i,t}}}$, where α is a positive constant no less than 1.5.

CC-UCB adopts the OFU principle to construct an upper bound of the ratio $\frac{\theta_i}{c_i}$. The main technical difficulty and correspondingly our novel contribution, however, lies in the theoretical analysis. This is because we have to deal with two types of regret: the regret caused by pulling "bad" arms ($\theta_i < c_i$); and that caused by pulling "good" arms in a wrong order. To the authors' best knowledge, the latter component has not been addressed in the bandit literature before, and is technically challenging. The overall regret analysis of CC-UCB is thus complicated and non-trivial.

We have the following main result for the cumulative regret upper bound of Algorithm 1.

Theorem 2: Denote $\Delta_i := c_i - \theta_i$. When $\alpha \geq 1.5$, the cumulative regret under Algorithm 1 is upper bounded as follows:

$$R(T) \leq \sum_{i\in[K]\backslash I^*} c_i \frac{16\alpha \log T}{\Delta_i^2} + O(1),$$

where $[K]\backslash I^*$ includes all items in $[K]$ except those in I^*. The proof of Theorem 2 is deferred to Section III-C.

Remark: In Theorem 2, the upper bound depends on $(c_i - \theta_i)^2$, while conventional upper bounds for UCB algorithms usually depend on the gap between the ranking parameters. It is because the regret caused by pulling "good" arms in a wrong order is bounded, as shown in Section III-C; Thus, the regret is mainly due to pulling "bad" arms, which is determined by arm i itself ($U_{i,t} > L_{i,t}$) and not related to the θ/c gap between the best arm and arm i. When c_is are *known* a priori, the upper bound can be reduced by a factor of 4.

Algorithm 1: Cost-aware Cascading UCB (CC-UCB).

```
Input: $\epsilon$, $\alpha$.
Initialization: Pull all arms in $[K]$ once, and observe
  their states and costs.
while $t$ do
    for $i = 1 : K$ do
        $U_{i,t} = \hat{\theta}_{i,t} + u_{i,t}$;
        $L_{i,t} = \max(\hat{c}_{i,t} - u_{i,t}, \epsilon)$;
        if $U_{i,t}/L_{i,t} > 1$ then $i \to I_t$;
        end if
    end for
    Rank arms in $I_t$ in the descending order of $\frac{U_{i,t}}{L_{i,t}}$.
    for $i = 1 : |I_t|$ do
        Pull arm $I_t(i)$ and observe $X_{I_t(i),t}, Y_{I_t(i),t}$;
        $i \to \tilde{I}_t$;
        if $X_{I_t(i),t} = 1$ break;
        end if
    end for
    Update $N_{i,t}$, $\hat{\theta}_{i,t}$, $\hat{c}_{i,t}$ for all $i \in \tilde{I}_t$.
    $t = t + 1$;
end while
```

C. Analysis of the Upper Bound

Define $\mathcal{E}_t := \{\exists i \in [K], |\hat{\theta}_{i,t} - \theta_i| > u_{i,t}$ or $|\hat{c}_{i,t} - c_i| > u_{i,t}\}$, i.e. there exists an arm whose sample average of reward or cost lies outside the corresponding confidence interval. Denote $\bar{\mathcal{E}}_t$ as the complement of $\mathcal{E}_t$. Then, we have the following observations.

Lemma 1: If $\mathbb{1}(\bar{\mathcal{E}}_t) = 1$, then, under Algorithm 1, all arms in I^* will be included in I_t.

Proof: According to Algorithm 1, arm i will be included in I_t if $\frac{U_{i,t}}{L_{i,t}} \geq 1$. When $\mathbb{1}(\mathcal{E}_t) = 0$, we have $|\hat{\theta}_{i,t} - \theta_i| < u_{i,t}$, $|\hat{c}_{i,t} - c_i| < u_{i,t}$. Thus, $\hat{\theta}_{i,t} + u_{i,t} \geq \theta_i$, $\hat{c}_{i,t} - u_{i,t} \leq \max\{\hat{c}_{i,t} - u_{i,t}, \epsilon\} \leq c_i$, which implies that $\frac{U_{i,t}}{L_{i,t}} \geq 1$. ■

Lemma 2: Under Algorithm 1, when $\alpha \geq 1.5$, we have $\sum_{t=1}^{T}\mathbb{E}[\mathbb{1}(\mathcal{E}_t)] \leq \psi := K(1 + \frac{4\pi^2}{3})$.

Proof:

$$\sum_{t=1}^{T}\mathbb{E}[\mathbb{1}(\mathcal{E}_t)]$$

$$\leq K + \sum_{t=K+1}^{T}\sum_{k\in[K]}\Big(\mathbb{P}\left[|\hat{\theta}_{k,t} - \theta_k| > u_{k,t}\right] + \mathbb{P}\left[|\hat{c}_{k,t} - c_k| > u_{k,t}\right]\Big) \quad (3)$$

$$= \sum_{k\in[K]} \sum_{t=K+1}^{T} \sum_{n=1}^{t} \left(\mathbb{P}\left[|\hat{\theta}_{k,t} - \theta_k| > \sqrt{\frac{\alpha \log t}{N_{k,t}}}, N_{k,t} = n \right] \right.$$

$$\left. + \mathbb{P}\left[|\hat{c}_{k,t} - c_k| > \sqrt{\frac{\alpha \log t}{N_{k,t}}}, N_{k,t} = n \right] \right) + K$$

$$\le K + \sum_{k\in[K]} \sum_{t=K+1}^{T} \sum_{n=1}^{t} 4 \exp\left(-2 \frac{\alpha \log t}{n} n \right) \quad (4)$$

$$= K + 4 \sum_{k\in[K]} \sum_{t=K+1}^{T} t^{-2\alpha+1} \le K + K \frac{4\pi^2}{3} = \psi, \quad (5)$$

where (4) follows from the Hoeffding's inequality. ■

Then, define $\mathcal{B}_t := \left\{ \exists i^*, j^* \in I^*, i < j, \text{ s.t. } \frac{U_{i^*,t}}{L_{i^*,t}} < \frac{U_{j^*,t}}{L_{j^*,t}} \right\}$, which represents the event that arms from I^* are not ranked in the correct order. Since those arms are pulled linearly often in order to achieve small regret, intuitively, $\mathcal{B}_t$ happens with small probability.

Lemma 3: Under Algorithm 1, when $\alpha \ge 1.5$, $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}(\mathcal{B}_t) \right] \le \zeta$, where ζ is a constant depending on K, α and $\{\theta_i, c_i\}_{i\in[K]}$.

The proof of Lemma 3 is derived based on the observation that the arms in I^* are pulled linearly often if $\bar{\mathcal{E}}_t$ is true, and the corresponding confidence intervals $u_{i,t}$ shrink fast in time. As $u_{i,t}$ decreases below a certain threshold, $\bar{\mathcal{B}}_t$ happens. The detailed proof of Lemma 3 can be found in Appendix B.

Lemma 4: Consider an ordered list I_t that includes all arms from I^* with the same relative order as in I^*. Then, under Algorithm 1, $\mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t] \le \sum_{i\in\tilde{I}_t\backslash I^*} c_i$.

Proof: First, we point out the difference between I_t and I^* is that in I_t, there may exist arms from $[K]\backslash I^*$ that are inserted between the arms in I^*. Denoted such ordered subset of arms as $I_t\backslash I^*$. Then, depending on the realization of the states of the arm on I_t, a random subset of $I_t\backslash I^*$ will be pulled (i.e., $\tilde{I}_t\backslash I^*$), resulting in a different regret. Denote the index of the last pulled arm in $\tilde{I}_t\backslash I^*$ as $\tilde{i}$. Then, depending on the state of arm $\tilde{i}$, there are two possible cases:

i) $X_{\tilde{i},t} = 0$. For this case, the regret comes from the cost of pulling the arms in $\tilde{I}_t\backslash I^*$ only. This is because if I^* were the list of arms to pull, with the same realization of arm states, the learner would only pull the arms in $\tilde{I}_t \cap I^*$ and receive the same reward. Thus, given $\tilde{I}_t$ and $X_{\tilde{i},t} = 0$, we have $\mathbb{E}[r_t^* - r_t \mid \tilde{I}_t, X_{\tilde{i},t} = 0] = \sum_{i\in\tilde{I}_t\backslash I^*} c_i$.

ii) $X_{\tilde{i},t} = 1$. This indicates that $\tilde{i}$ is the last arm on $\tilde{I}_t$ due to the stopping condition. For this case, the learner spends costs on pulling the arms in $\tilde{I}_t\backslash I^*$ but also receives the full reward one. If I^* were the list of arms to pull, with the same realization of arm states, the learner would first pull all arms in $\tilde{I}_t \cap I^*$. Since the states of such arms should be 0, she would then continue pulling the remaining arms in I^* if there is any. Denote the net reward obtained from the remaining pullings as $r(I^*\backslash\tilde{I}_t)$. Then, $\mathbb{E}[r(I^*\backslash\tilde{I}_t)] \le 1$. Therefore, given $\tilde{I}_t$ and $X_{\tilde{i},t} = 1$, we have $\mathbb{E}[r_t^* - r_t | \tilde{I}_t, X_{\tilde{i},t} = 1] = \mathbb{E}[r(I^*\backslash\tilde{I}_t)] - (1 - \sum_{i\in\tilde{I}_t\backslash I^*} c_i) \le \sum_{i\in\tilde{I}_t\backslash I^*} c_i$.

Combining both cases, we have $\mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t] = \mathbb{E}[r_t^* - r_t \mid \tilde{I}_t] \le \sum_{i\in\tilde{I}_t\backslash I^*} c_i$, which completes the proof. ■

Next, we consider the regret resulted from including arms outside I^* in the list I_t. We focus on the scenario when both $\mathcal{E}_t$ and $\mathcal{B}_t$ are false. Notice that the condition of Lemma 4 is satisfied in this scenario. Thus,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}(\bar{\mathcal{B}}_t) \mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t] \right]$$

$$\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}(\bar{\mathcal{B}}_t) \left(\sum_{i\in\tilde{I}_t\backslash I^*} c_i \right) \right] \quad (6)$$

$$= \mathbb{E}\left[\sum_{i\in[K]\backslash I^*} \sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}(\bar{\mathcal{B}}_t) \mathbb{1}(i \in \tilde{I}_t) c_i \right] \quad (7)$$

$$\le \mathbb{E}\left[\sum_{i\in[K]\backslash I^*} \sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}\left(\frac{U_{i,t}}{L_{i,t}} > 1 \right) \mathbb{1}(i \in \tilde{I}_t) c_i \right] \quad (8)$$

$$\le \sum_{i\in[K]\backslash I^*} c_i \frac{16\alpha \log T}{\Delta_i^2}, \quad (9)$$

where (6) is based on Lemma 4; (9) is due to the fact that $\mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\frac{U_{i,t}}{L_{i,t}} > 1)$ is always less than or equal to $\mathbb{1}(\theta_i + 2u_{i,t} > c_i - 2u_{i,t})$, which is equivalent to $\mathbb{1}(N_{i,t} < \frac{16\alpha\log t}{\Delta_i^2})$.

Denote δ^* as the largest possible per-step regret, which is bounded by $\sum_{i\in[K]} c_i$, corresponding to the worst case scenario that the learner pulls all arms but does not receive reward one. Then, combining the results from above, we have

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} [\mathbb{1}(\mathcal{E}_t) + \mathbb{1}(\bar{\mathcal{E}}_t)] \mathbf{reg}_t \right]$$

$$\le \delta^* \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{E}_t) + \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\mathcal{B}_t) \right]$$

$$+ \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}(\bar{\mathcal{B}}_t) \mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t] \right]$$

$$\le \delta^*(\zeta + \psi) + \sum_{j\in[K]\backslash I^*} c_j \frac{16\alpha \log T}{\Delta_j^2},$$

which proves Theorem 2.

D. Lower Bound

Before presenting the lower bound, we first define α-consistent policies.

Definition 1: Consider online policies that sequentially pull arms in I_t until one arm with state 1 is observed or all arms are pulled. If $\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(I_t \neq I^*)] = o(T^\alpha)$ for any $\alpha \in (0, 1)$, the policy is α-consistent.

Lemma 5: For any ordered list I_t, the per-step regret in step t is lower bounded by $\mathbb{E}[\mathbf{reg}_t] \ge \mathbb{E}[\sum_{i\in\tilde{I}_t\backslash I^*} (c_i - \theta_i)]$.

Proof: Consider an ordered list $I_t^* := I_t \cap I^*$, i.e., the sublist of I_t that only contains the arms from I^* while keeping their relative order in I_t. Denote $r_t(I_t^*)$ as the reward obtained at step t by pulling the arms in I_t^* sequentially or until all arms are pulled. We have

$$\begin{aligned}\mathbb{E}[\mathbf{reg}_t] &= \mathbb{E}[r_t^* - r_t(I_t^*) + r_t(I_t^*) - r_t] \\ &\geq \mathbb{E}[r_t(I_t^*) - r_t], \qquad (10)\end{aligned}$$

where inequality (10) follows from the fact that I^* maximizes the expected reward in every step.

Similar to the proof of Lemma 4, we denote the index of the last pulled arm in $\tilde{I}_t \backslash I^*$ as $\tilde{i}$. Then, given $\tilde{I}_t$ and $X_{\tilde{i},t} = 0$, we have $\mathbb{E}[r_t(I_t^*) - r_t \mid \tilde{I}_t, X_{\tilde{i},t} = 0] = \sum_{i \in \tilde{I}_t \backslash I^*} c_i$.

If $X_{\tilde{i},t} = 1$, based on the assumption that all policies should stop pulling in a step if a state 1 is observed, we infer that the arms in $\tilde{I}_t \cap I^*$ should have state 0. If I_t^* were the list of arms to pull, with the same realization of arm states, the learner would continue pulling the remaining arms in I_t^* if there is any. Denote the net reward obtained from the rest pulling as $r(I_t^* \backslash \tilde{I}_t)$. Then, due to the definition of I^*, we have $\mathbb{E}[r(I_t^* \backslash \tilde{I}_t)] \geq 0$. Therefore, given $\tilde{I}_t$ and $X_{\tilde{i},t} = 1$, we have $\mathbb{E}[r_t(I_t^*) - r_t \mid \tilde{I}_t, X_{\tilde{i},t} = 1] = \mathbb{E}[r(I_t^* \backslash \tilde{I}_t)] - \left(1 - \sum_{i \in \tilde{I}_t \backslash I^*} c_i\right) \geq \sum_{i \in \tilde{I}_t \backslash I^*} c_i - 1$.

Combining both cases, we have $\mathbb{E}[r_t^* - r_t \mid \tilde{I}_t] \geq \sum_{i \in \tilde{I}_t \backslash I^*} c_i - \theta_{\tilde{i}} \geq \sum_{i \in \tilde{I}_t \backslash I^*} (c_i - \theta_i)$. Taking expectation with respect to $\tilde{I}_t$, we obtain the lower bound for $\mathbb{E}[\mathbf{reg}_t]$. ■

Theorem 3: Under any α-consistent policy, we have

$$\liminf_{T \to \infty} \frac{R(T)}{\log T} \geq \sum_{i \in [K] \backslash I^*} \frac{c_i - \theta_i}{d(\theta_i; c_i)},$$

where $d(\theta_i; c_i)$ is the KL divergence of Bernoulli distributions with means θ_i and c_i.

Proof: According to Lemma 5, we have

$$\mathbb{E}[\mathbf{reg}_t] \geq \mathbb{E}\left[\sum_{i \in [K] \backslash I^*} \mathbb{1}(i \in \tilde{I}_t)(c_i - \theta_i)\right]. \qquad (11)$$

Therefore,

$$\begin{aligned}R(T) &\geq \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i \in [K] \backslash I^*} \mathbb{1}(i \in \tilde{I}_t)(c_i - \theta_i)\right)\right] \qquad (12) \\ &= \mathbb{E}\left[\sum_{i \in [K] \backslash I^*} (c_i - \theta_i)\left(\sum_{t=1}^{T} \mathbb{1}(i \in \tilde{I}_t)\right)\right] \qquad (13) \\ &= \sum_{i \in [K] \backslash I^*} (c_i - \theta_i)\mathbb{E}[N_{i,T}]. \qquad (14)\end{aligned}$$

Next, we aim to provide a lower bound for $\mathbb{E}[N_{i,T}]$. Denote ν as the original distribution of the arm rewards and costs with mean $\{\theta_i, c_i\}_{i=1}^{K}$. Without loss of generality, we assume $\frac{\theta_1}{c_1} \geq \ldots \frac{\theta_L}{c_L} > 1$. Construct an alternative distribution ν' where we change the expected reward of arm $i \in [K] \backslash I^*$ from θ_i to θ_i', such that $\frac{\theta_i'}{c_i} \in (1, \frac{\theta_L}{c_L})$. We note that under distribution ν', arm i becomes the additional good arm that is appended to the original optimal list $[L]$.

Based on the definition of the α-consistent policy, for any good arm $k \in [L]$, we have

$$\begin{aligned}&\mathbb{E}_\nu\left[N_{k,T}\right] = \mathbb{E}_{\nu'}\left[N_{k,T}\right] \\ &= \mathbb{E}_\nu\left[\sum_{t=1}^{T} \mathbb{1}\{k \in \tilde{I}_t, I_t = I^*\}\right] + \mathbb{E}_\nu\left[\sum_{t=1}^{T} \mathbb{1}\{k \in \tilde{I}_t, I_t \neq I^*\}\right] \\ &= \Pi_{j=1}^{k-1}(1 - \theta_j)\mathbb{E}_\nu\left[\sum_{t=1}^{T} \mathbb{1}\{I_t = I^*\}\right] + o(T^\alpha) \forall \alpha \in (0,1) \\ &= \Pi_{j=1}^{k-1}(1 - \theta_j)T + o(T^\alpha), \forall \alpha \in (0,1), \qquad (15)\end{aligned}$$

where the first equality holds because appending an extra good arm to $[L]$ would not change the behavior of the arms in $[L]$. Then, for arm i, we have

$$\mathbb{E}_{\nu'}\left[N_{i,T}\right] = \Pi_{j=1}^{L}(1 - \theta_j)T + o(T^\alpha), \forall \alpha \in (0,1).$$

Thus, under any α-consistent policy, we have

$$\mathbb{E}_{\nu'}[\Pi_{j=1}^{L}(1 - \theta_j)T - N_{i,T}] = o(T^\alpha).$$

Following the approach in [29], by letting θ_i approach c_i, we have

$$\lim_{T \to \infty} \mathbb{P}_\nu\left[N_{i,T} < \frac{\log T}{d(\theta_i; c_i)}\right] = 0.$$

Thus,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\nu[N_{i,T}]}{\log T} \geq \frac{1}{d(\theta_i; c_i)}.$$

Combining with (14), we have the lower bound hold. ■

Remark: We note that $d(\theta_i, c_i) \leq \frac{(c_i - \theta_i)^2}{c_i(1 - c_i)}$ [30]. Thus, $\frac{c_i - \theta_i}{d(\theta_i, c_i)} \geq \frac{(c_i - \theta_i)c_i(1 - c_i)}{(c_i - \theta_i)^2} = \frac{c_i(1 - c_i)}{\Delta_i}$. It differs from the coefficient of the $\log(T)$ term associated with arm i in Theorem 2 by a factor of $(1 - c_i)\Delta_i$. Although the cascading bandit feedback structure leads to the mismatch between the coefficients, we note that both the upper and lower bounds only involve the "bad" arms in $[K] \backslash I^*$, and have the same scaling in T. Thus, we conclude that Algorithm 1 achieves order-optimal regret performance.

IV. Known Immediate Costs

In this section, we consider a scenario where the cost of pulling each arm (i.e., $Y_{i,t}$) is available to the learner at the beginning of step t, where $Y_{i,t}$ evolves according to a *possibly unknown* i.i.d. distribution. Without a priori statistics about $\{X_{i,t}\}$, the learner needs to decide I_t based on previous observations $\cup_{\tau=1}^{t-1}\{X_{i,\tau}\}_{i \in \tilde{I}_\tau}$ and immediate costs $\{Y_{i,t}\}_{i \in [K]}$, and $\tilde{I}_t$ based on observed states in step t, so that the cumulative regret $R(T)$ is minimized.

Due to the availability of the immediate costs, the optimal offline policy is no longer fixed but varying according to $\{Y_{i,t}\}_{i \in [K]}$. To see this, we first make the following assumptions:

Assumptions 2: 1) There exists a constant $\tilde{\epsilon} > 0$, such that $|\theta_i - Y_{i,t}| \geq \tilde{\epsilon}$ with probability 1 for all $i \in [K]$. 2) The support of $Y_{i,t}$ is $[l_i, h_i]$, where $0 < l_i \leq h_i < \infty$ for all $i \in [K]$. 3) With probability 1, $|\frac{\theta_i}{Y_{i,t}} - \frac{\theta_j}{Y_{j,t}}| \geq \Delta > 0$ for all $i, j \in [K], i \neq j$.

Then, the optimal offline policy at step t is as follows.

Theorem 4: Assume $\{Y_{i,t}\}_{i\in[K]}$ is given at the beginning of step t. Arrange the arm indices such that

$$\frac{\theta_{1^*}}{Y_{1^*,t}} \geq \cdots \geq \frac{\theta_{L_t^*}}{Y_{L_t^*,t}} > 1 > \frac{\theta_{(L_t+1)^*}}{Y_{(L_t+1)^*,t}} \geq \cdots \geq \frac{\theta_{K^*}}{Y_{K^*,t}}.$$

Then, the optimal list of arms to pull at time t is $I_t^* = \{1^*, 2^*, \ldots, L_t^*\}$, and the corresponding optimal expected net reward at step t is $\mathbb{E}[r_t^* | \{Y_{i,t}\}_{i\in[K]}] = \sum_{i=1}^{L_t}(\theta_{i^*} - Y_{i^*,t})\prod_{j=1}^{i-1}(1-\theta_{j^*})$.

The proof of Theorem 4 follows a similar approach to the proof of Theorem 1, thus is omitted for the brevity of the paper.

Remark: Compared with the unknown cost setting in Section IV, one critical difference under the known cost setting and is that, the optimal list to pull each time is not fixed but *context-dependent*, where the context is the instantaneous costs of arms, i.e., I_t^* varies according to $\{Y_{i,t}\}_{i\in[K]}$. Denote $\mathbf{Y}_t := [Y_{1,t}, \ldots, Y_{K,t}]$. Then, we partition the range of $\mathbf{Y}_t$ into subsets $\mathcal{Y}_1, \ldots, \mathcal{Y}_M$, where for any $\mathbf{Y}_t$ lying in the same subset, the corresponding optimal list to pull is the same. For finite number of arms, the total number of different optimal lists (i.e., M) must be finite as well. Define $\rho_m := \mathbb{P}[\mathbf{Y}_t \in \mathcal{Y}_m]$, $m = 1, 2, \ldots, M$, and the corresponding optimal list as I^m. We point out that I^m could be an empty list, i.e., no arm should be pulled under the corresponding cost vectors. Besides, we partition the arms into two subsets, depending on whether they are included in the optimal lists $\{I^m\}_m$. Specifically, we use $\mathcal{S}_1$ to denote the arms that are included in at least one of the optimal lists with a non-zero probability, and use $\mathcal{S}_2$ to denote its complement, i.e., $\mathcal{S}_2 := [K]\backslash\mathcal{S}_1$. Intuitively, in order to achieve small learning regret, all arms in $\mathcal{S}_1$ should be pulled frequently in time, and the regret thus mainly depends on $\mathcal{S}_2$.

Assumption 2.1 implies that for any arm $i \in \mathcal{S}_1$, it can generate a positive net-reward with probability 1 when it is pulled as an optimal arm in I_t^*. Assumption 2.3 ensures that with probability 1, there is a non-zero net reward margin for all arms included in I_t^*.

A. Online Algorithm and Upper Bound

Motivated by the optimal offline policy, we modify Algorithm 1 slightly by replacing the LCB of $\hat{c}_{i,t}$ (i.e., $L_{i,t}$) with $Y_{i,t}$. Denote this modified version as CC-UCB2. We have the following result for the corresponding cumulative regret upper bound.

Theorem 5: Denote $\underline{\Delta}_i := l_i - \theta_i$. Then, when $\alpha \geq 1.5$, the cumulative regret under CC-UCB2 is upper bounded as follows:

$$R(T) \leq \sum_{i\in\mathcal{S}_2} h_i \frac{4\alpha \log T}{\underline{\Delta}_i^2} + O(1).$$

Remark: We note that for $i \in \mathcal{S}_2$, we must have $Y_{i,t} > \theta_i$ with probability 1, which implies that $\theta_i < l_i$ under Assumption 2.1. As we expect, the upper bound mainly depends on $\mathcal{S}_2$. If $\mathcal{S}_2 = \emptyset$, the expected regret can be bounded by a *constant*. This is in stark contrast with the unknown cost case discussed in Section III. Theorem 1 indicates that the upper bound depends on the arms outside I^*, which is decided by the mean of $Y_{i,t}$. With the instantaneous cost information $\mathbf{Y}_t$ available, the learner is able to utilize the information to adaptively decide the list to pull each time. Therefore, when the instantaneous cost of pulling one arm is low, it has a higher chance to be ranked higher on the list, and to be actually pulled afterwards. The i.i.d. assumptions on $Y_{i,t}$ enables the learner to perform such *opportunistic exploration* [31] frequently, thus reducing the regret caused by exploration.

B. Analysis of the Upper Bound

In order to analyze the performance of CC-UCB2, we modify the definition of $\mathcal{E}_t$ by focusing on the sample average of rewards only. Then, Lemma 1 and Lemma 2 still hold.

We also modify the definition of $\mathcal{B}_t$ by replacing $L_{i,t}$ with $Y_{i,t}$ and replacing I^* with I_t^*. Similar to Lemma 3, we have the following observation.

Lemma 6: Under CC-UCB2, when $\alpha \geq 1.5$, we have $\mathbb{E}[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\mathcal{B}_t)\mathbb{1}(\mathbf{Y}_t \in \mathcal{Y}_m)] \leq \zeta^m$, $\forall m$, where ζ^m is a constant depending on $K, \alpha, \{\theta_i, l_i\}_{i\in[K]}$.

The proof is deferred to Appendix C.

Besides, Lemma 4 can be extended straightforwardly as follows.

Lemma 7: Consider an ordered list I_t that includes all arms from I_t^* with the same relative order as in I_t^*. Then, under CC-UCB2, $\mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t, \mathbf{Y}_t] \leq \sum_{i\in\tilde{I}_t\backslash I_t^*} Y_{i,t}$.

With Lemma 7, we aim to bound $\mathbb{E}[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t)\mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t, \mathbf{Y}_t]]$. We note that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t)\mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t, \mathbf{Y}_t]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t) \sum_{i\in\tilde{I}_t\backslash I_t^*} Y_{i,t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t)\left(\sum_{i\in\mathcal{S}_2\cap\tilde{I}_t} Y_{i,t} + \sum_{i\in\tilde{I}_t\backslash(\mathcal{S}_2\cup I_t^*)} Y_{i,t}\right)\right].
\end{aligned}$$

Following similar steps after (9), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t) \sum_{i\in\mathcal{S}_2\cap\tilde{I}_t} Y_{i,t}\right] \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{S}_2}\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(\frac{U_{i,t}}{Y_{i,t}} > 1\right)\mathbb{1}(i \in \tilde{I}_t)Y_{i,t}\right] \\
&\leq \mathbb{E}\left[\sum_{i\in\mathcal{S}_2}\sum_{t=1}^T \mathbb{1}\left(N_{i,t} < \frac{4\alpha\log t}{\underline{\Delta}_i^2}\right)\mathbb{1}(i \in \tilde{I}_t)h_i\right] \\
&\leq \sum_{i\in\mathcal{S}_2} h_i \frac{4\alpha \log T}{\underline{\Delta}_i^2}.
\end{aligned}$$

Hence, the main difference between the upper bound analysis in this setting and that in Section III, is the regret generated when arms in $\mathcal{S}_1$ are pulled as a "bad" arm under certain cost vectors $\mathbf{Y}_t$. We aim to show that this part can be bounded by a constant. The intuition is that arms in $\mathcal{S}_1$ should be pulled as a "good" arm frequently, and their reward statistics can be estimated accurately. Thus, the probability that they are included in I_t as a bad arm diminishes quickly in time.

Denote $l_i^m := \min\{Y_{i,t}|\mathbf{Y}_t \in \mathcal{Y}_m\}$. Then, according to Theorem 4, if $\mathbf{Y}_t \in \mathcal{Y}_m$, for all $i \in \mathcal{S}_1\backslash I^m$, we must have $Y_{i,t} \ge l_i^m > \theta_i$ with probability 1. Denote $\Delta_i^m := l_i^m - \theta_i$, and $\tilde{\Delta}_i := \min_{m:i\notin I^m} \Delta_i^m$. Then, we have

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t) \sum_{i\in\tilde{I}_t\backslash(\mathcal{S}_2\cup I_t^*)} Y_{i,t}\right]$$

$$\le \sum_{m=1}^M \mathbb{E}\Bigg[\sum_{i\in\mathcal{S}_1\backslash I^m}\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(\frac{U_{i,t}}{Y_{i,t}} > 1\right)\mathbb{1}(i\in\tilde{I}_t) \cdot Y_{i,t}\mathbb{1}(\mathbf{Y}_t\in\mathcal{Y}_m)\Bigg]$$

$$\le \sum_{m=1}^M \mathbb{E}\Bigg[\sum_{i\in\mathcal{S}_1\backslash I^m}\sum_{t=1}^T \mathbb{1}\left(N_{i,t} < \frac{4\alpha\log t}{\tilde{\Delta}_i^2}\right)\mathbb{1}(i\in\tilde{I}_t) \cdot \mathbb{1}(\mathbf{Y}_t\in\mathcal{Y}_m)h_i\Bigg] \quad (16)$$

$$\le \mathbb{E}\left[\sum_{i\in\mathcal{S}_1}\sum_{t=1}^T \mathbb{1}\left(N_{i,t} < \frac{4\alpha\log t}{\tilde{\Delta}_i^2}\right)h_i\right]. \quad (17)$$

Therefore, it suffices to show that (17) can be bounded by a constant.

Let $N_{i,t}^m$ be the number of times that arm i is pulled before step t under a cost vector in $\mathcal{Y}_m$, and $\hat{N}_{i,t}^m$ be the number of times that arm i is pulled between step $\lceil t/2\rceil$ and t under a cost vector in $\mathcal{Y}_m$. We have the following lemma.

Lemma 8: For any $i \in \mathcal{S}_1$, any m s.t arm $i \in I^m$, when $t > 2K$, we have

$$\mathbb{P}\left[\hat{N}_{i,t}^m < \frac{\rho_m p_i t}{4}\right] \le 2K\left(\frac{t}{2}+1\right)\left(\frac{t}{2}\right)^{-2\alpha+1} + \exp\left(-\frac{\rho_m^2 p_i^2}{16}t\right) := C_{i,t}^m.$$

where $p_i := \frac{\prod_{j=1}^K(1-\theta_j)}{(1-\theta_i)}$.

The proof of Lemma 8 is deferred to Appendix D.

For any $i \in \mathcal{S}_1$, let $\hat{m} := \arg\max_{m:i\in I^m}\rho_m$. Besides, we let n_i be $\frac{32\alpha}{\rho_{\hat{m}}p_i\tilde{\Delta}_i^2}\log\frac{32\alpha}{\rho_{\hat{m}}p_i\tilde{\Delta}_i^2}$ if $\frac{32\alpha}{\rho_{\hat{m}}p_i\tilde{\Delta}_i^2} \ge e$ and let it be 2 otherwise. Then, following the approach in the proof of Lemma 3, we can show that for any $t \ge n_i$, we must have $\frac{\rho_{\hat{m}}p_i t}{4} \ge \frac{4\alpha\log t}{\tilde{\Delta}_i^2}$. Since $N_{i,t} \ge N_{i,t}^{\hat{m}} \ge \hat{N}_{i,t}^{\hat{m}}$, we have

$$(17) \le \mathbb{E}\left[\sum_{i\in\mathcal{S}_1}\left(n_i + \sum_{t=n_i}^T \mathbb{1}\left(\hat{N}_{i,t}^{\hat{m}} < \frac{\rho_{\hat{m}}p_i t}{4}\right)\right)h_i\right]$$

$$= \sum_{i\in\mathcal{S}_1}\left(n_i + \sum_{t=n_i}^T C_{i,t}^{\hat{m}}\right)h_i$$

$$\le \sum_{i\in\mathcal{S}_1}\left(n_i + \sum_{t=1}^\infty C_{i,t}^{\hat{m}}\right)h_i := \xi. \quad (18)$$

Combining Lemma 6, (16) and (18), we have

$$R(T) \le \delta^*\sum_{t=1}^T \mathbb{E}\left[\mathbb{1}(\mathcal{E}_t) + \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\mathcal{B}_t)\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t)\mathbb{E}[\mathbf{reg}_t \mid \tilde{I}_t, \mathbf{Y}_t]\right] \quad (19)$$

$$\le \sum_{i\in\mathcal{S}_2} h_i\frac{4\alpha\log T}{\Delta_i^2} + \delta^*\left(\psi + \xi + \sum_{m=1}^M \zeta^m\right), \quad (20)$$

which completes the proof of Theorem 5.

C. Lower Bound

Following the proof of Lemma 5, we have the following observations.

Lemma 9: For any ordered list I_t, given the pulled list $\tilde{I}_t$ and $\mathbf{Y}_t$, the per-step regret in step t is lower bounded by $\mathbb{E}[\mathbf{reg}_t|\tilde{I}_t, \mathbf{Y}_t] \ge \mathbb{E}[\sum_{i\in\tilde{I}_t\backslash I_t^*}(Y_{i,t} - \theta_i)]$.

Theorem 6: Under any α-consistent policy, we have

$$\liminf_{T\to\infty}\frac{R(T)}{\log T} \ge \sum_{i\in\mathcal{S}_2}\frac{l_i - \theta_i}{d(\theta_i; l_i)},$$

where $d(\theta_i; l_i)$ is the KL divergence of Bernoulli distributions with means θ_i and l_i.

Proof: According to Lemma 5, we have

$$\mathbb{E}[\mathbf{reg}_t] \ge \mathbb{E}\left[\sum_{i\in[K]\backslash I_t^*}\mathbb{1}(i\in\tilde{I}_t)(Y_{i,t} - \theta_i)\right] \ge \mathbb{E}\left[\sum_{i\in\mathcal{S}_2}\mathbb{1}(i\in\tilde{I}_t)(l_i - \theta_i)\right].$$

Therefore,

$$R(T) \ge \mathbb{E}\left[\sum_{i\in\mathcal{S}_2}(l_i - \theta_i)\left(\sum_{t=1}^T \mathbb{1}(i\in\tilde{I}_t)\right)\right] \quad (21)$$

$$= \sum_{i\in\mathcal{S}_2}(l_i - \theta_i)\mathbb{E}[N_{i,T}]. \quad (22)$$

Recall that for $i \in \mathcal{S}_2$, $\theta_i < l_i$. Following a similar approach as in the proof of Theorem 3, we have

$$\liminf_{T\to\infty}\frac{\mathbb{E}[N_{i,T}]}{\log T} \ge \frac{1}{d(\theta_i; l_i)}. \quad (23)$$

Combining (22) with (23), we obtain the lower bound. ■

Comparing Theorem 5 with Theorem 6, we conclude that Algorithm CC-UCB2 is order-optimal.

V. Extension: Fail-Safe CCB

In this section, we discuss a possible extension of the CCB model by considering an extra cost for the cascading failure in each step. Specifically, we assume the learner will sequentially check the recommended list I_t until an arm with state 1 is observed; if it checks the whole list and finds no arm with state 1, an additional positive *fixed* cost C would be incurred. Note that this is in addition to the random cost $Y_{i,t}$ incurred for each examined item in the list.

With the inclusion of a failure penalty C, the net reward at each step can be expressed as

$$r_t = 1 - \prod_{i=1}^{|\tilde{I}_t|}(1 - X_{\tilde{I}_t(i),t}) - \sum_{i=1}^{|\tilde{I}_t|} Y_{\tilde{I}_t(i),t} - C\prod_{i=1}^{|I_t|}(1 - X_{I_t(i),t}). \tag{24}$$

In order to solve the problem, we start with the optimal offline policy. Similar to Theorem 1, we observe that all arms on I_t should be ranked in the descending order of θ_i/c_i. Besides, the optimal offline policy must contain all arms i with $\theta_i > c_i$. This can be proved in a way similar to the proof of Theorem 1 in Appendix A.

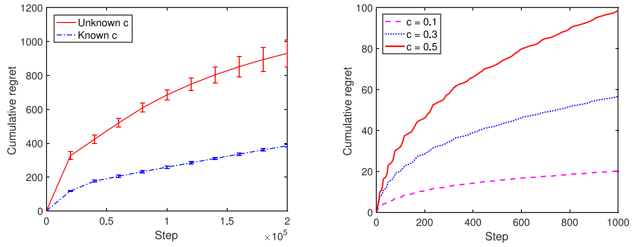
The question is whether we should add arms with $\theta_i \le c_i$ to the list. Consider two ordered lists $I = (I(1), I(2), \ldots, I(|I|))$ and $I' = (I(1), I(2), \ldots, I(|I|), k)$, where the only difference between them is the extra arm $k \in [K]\backslash I$ in I'. Then, the difference between the expected net rewards of I and I' is

$$\begin{aligned}
&\mathbb{E}[r(I')] - \mathbb{E}[r(I)] \\
&= (\theta_k - c_k)\prod_{i=1}^{|I|}(1 - \theta_{I(i)}) - C\prod_{i=1}^{|I|}(1 - \theta_{I(i)})(1 - \theta_k) \\
&\quad + C\prod_{i=1}^{|I|}(1 - \theta_{I(i)}) \\
&= \prod_{i=1}^{|I|}(1 - \theta_{I(i)})\left[(1 + C)\theta_k - c_k\right].
\end{aligned} \tag{25}$$

As we can see, as long as $(1 + C)\theta_k > c_k$, adding k to the end of I strictly improves the expected net reward.

Combining the observations together, we can see that the optimal offline policy is to rank all arms in the descending order of θ_i/c_i and include those with the ratio above $\frac{1}{1+C}$. Compared with Theorem 1, more arms will be included if $C > 0$. This is intuitive as the learner would examine more arms in order to avoid the failure penalty.

Since the threshold structure preserves in the offline setting, CC-UCB and CC-UCB2 can be slightly modified to develop the optimal offline policy. Specifically, we just need to include the arms with $\frac{U_{i,t}}{L_{i,t}} > \frac{1}{1+C}$ in I_t under CC-UCB, and include those with $\frac{U_{i,t}}{Y_{i,t}} > \frac{1}{1+C}$ under CC-UCB2.



(a) Synthetic data (b) Real-world data

Fig. 1. Cumulative regret versus step under CC-UCB.

Regarding the regret of the online algorithms, we can show that under the known cost setting, the cumulative regret is upper bounded as follows:

$$R(T) \le \delta^*(\zeta + \psi) + \sum_{j \in [K]\backslash I^*} c_i \frac{4(2 + C)^2 \alpha \log T}{\bar{\Delta}_i^2},$$

where $\bar{\Delta}_i = c_i - (1 + C)\theta_i$. This is because changing the threshold on $\frac{U_{i,t}}{L_{i,t}}$ would only affect the upper bound in Eqn. (9).

For the corresponding lower bound, following the approach to proving Theorem 3, we can show that under any α-consistent policy,

$$\liminf_{T\to\infty} \frac{R(T)}{\log T} \ge \sum_{i \in [K]\backslash I^*} \frac{c_i - (1 + C)\theta_i}{d(\theta_i; \frac{c_i}{1+C})} \quad \text{almost surely.}$$

For the known-cost setting, the upper and lower bounds would follow similar forms as those in Theorem 5 and Theorem 6, and we omit them for the brevity of this paper.

VI. Simulation Results

In this section we resort to numerical experiments to evaluate the performances of the proposed CC-UCB and CC-UCB2 algorithms. We first consider the unknown immediate costs setting, and then investigate the known immediate costs setting. For both algorithms, we set $\alpha = 1.5$.

A. Unknown Immediate Costs

We first evaluate the performance of CC-UCB under the unknown cost setting. We set $\epsilon = 10^{-5}$. Both synthetic and real-world datasets are used.

1) Synthetic Data: We consider a 6-arm bandits setting with parameters $\boldsymbol{\theta} = \{0.8, 0.7, 0.6, 0.5, 0.4, 0.3\}$, and the costs follow an i.i.d. Bernoulli distribution with parameter $c = 0.55$ for all arms. According to the UCR-T1 policy, we have $L = 3$, i.e., the first three arms should be included in I^*. We then perform the CC-UCB algorithm under the assumption that both $\boldsymbol{\theta}$ and c are unknown to the learning agent. We run it for $T = 2 \times 10^5$ steps, and average the cumulative regret over 20 runs. The result is plotted in Fig. 1(a). The error bar corresponds to one standard deviation of the regrets in 20 runs. We also study the case where the mean of the cost c is known beforehand, however, the cost of each arm is still random. The result is also plotted in the same figure. As we observe, both curves increase sublinearly in T, which is consistent with the $O(\log T)$ bound we derive in

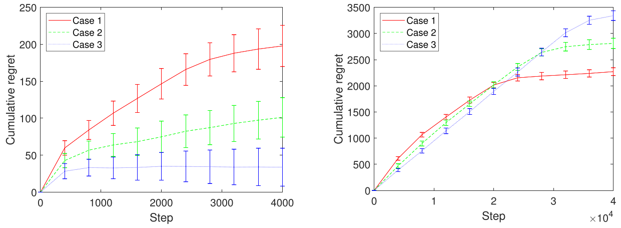
TABLE I
T-STEP REGRET IN $T = 10^5$ STEPS

K	L	Δ_{L+1}	Known c	Unknown c
6	1	0.1	580.3288	2.2862e+03
6	3	0.1	352.8772	1.4453e+03
6	5	0.1	117.5846	364.6771
12	1	0.1	2.5284e+03	1.0225e+04
12	3	0.1	1.2996e+03	4.8120e+03
12	5	0.1	387.8936	1.3728e+03
6	1	0.05	1.1536e+03	4.7941e+03
6	3	0.05	697.7550	1.4431e+03
6	5	0.05	160.7688	212.0552

Theorem 2. The regret with known cost statistics is significantly smaller than that of the unknown cost statistics case.

Next, we evaluate the impact of system parameters on the performance of the algorithm. We vary K and L, i.e., the total number of arms K, and the number of arms in I^*, respectively. We also change Δ_i, i.e., $c_i - \theta_i$, for $i \in [K]\backslash I^*$. Specifically, we set $\theta_i = 0.5$ for $i \in I^*$, $\theta_i = 0.3$ for $i \in [K]\backslash I^*$, and let c_i be a constant c across all arms. By setting Δ_{L+1}, the value of c can be determined. The cumulative regrets at $T = 10^5$ averaged over 20 runs are reported in Table I. We observe four major trends. First, the regret increases when the number of arms K doubles. Second, the regret decreases when the number of arms in I^* (i.e., L) increases. Third, a prori knowledge of cost statistics always improve the regret. Fourth, when c is known, the regret increases as Δ_{L+1} decreases. These trends are consistent with Theorem 2. However, the dependency on Δ_{L+1} when c is unknown is not obvious from the experiment results. This might be because the algorithm depends on the estimation of the cost as well as the arm state, and the complicated interplay between cost and the optimal arm pulling policy makes the dependency on Δ_{L+1} hard to discern.

2) Real-World Data: In this section, we test the proposed CC-UCB algorithm using real-world data extracted from the click log dataset of Yandex Challenge [32]. The click log contains complex user behaviors. To make it suitable for our algorithm, we extract data that contains click history of a specific query. We choose the total number of links for the query to be 15 and set a constant and known cost c for all of them. We estimate the probability that a user would click a link based on the dataset and use it as the ground truth θ_i. We then test the CC-UCB based on the structure of the optimal offline policy. We plot the cumulative regret in Fig. 1(b) with different values of c. We observe that the cumulative regret grows sub-linearly in time, and monotonically increases as c increases, which are consistent with Theorem 2. This indicates that the CC-UCB algorithm performs very well even when some of the assumptions (such as the i.i.d. evolution of arm states) we used to derive the performance bounds do not hold.



(a) (b)

Fig. 2. Cumulative regret versus step.

B. Known Immediate Costs

In this section, we consider the known immediate cost setting and investigate the performance of CC-UCB2. We fix the reward distributions of the bandits as in the previous setting, and modify the cost distributions as follows.

We consider a 5-arm bandits setting, where each arm may follow one of the following three parameter settings: a) $\theta_i = 0.5$, $l_i = 0.2$, $h_i = 0.4$; b) $\theta_i = 0.5$, $l_i = 0.4$, $h_i = 0.8$; c) $\theta_i = 0.3$, $l_i = 0.4$, $h_i = 0.6$. For all settings, we assume $\mathbb{P}[\mathcal{Y}_{i,t} = h_i] = \mathbb{P}[\mathcal{Y}_{i,t} = l_i] = 1/2$. Thus, if the immediate costs of arms are unknown to the learner before they are pulled, arms with parameter setting a) are good arms and will be included in I^*, while the arms with parameter settings b) and c) are bad arms and not included in I^*. On the other hand, if the immediate costs of arms are known before the decision making, arms with parameter setting a) will always be good arms, those with parameter setting b) will be good arms when $\mathcal{Y}_{i,t} = l_i$, and those with parameter setting c) will never be good arms, thus lying in $\mathcal{S}_2$.

We compare the performances of CC-UCB for three cases as follows:

- Case 1: One arm with parameter setting a), two arms with parameter setting b), and two arms with parameter setting c).
- Case 2: One arm with parameter setting a), three arms with parameter setting b), and one arm with parameter setting c).
- Case 3: One arm with parameter setting a), four arms with parameter setting b), and no arm with parameter setting c).

The cumulative regret over 200 sample paths under CC-UCB2 is plotted in Fig. 2(a). We note that the cumulative regret monotonically increases in $|\mathcal{S}_2|$ when T is sufficiently large. For cases 1 and 2, the regret increases sublinearly in T. As $|\mathcal{S}_2|$ becomes zero for case 2, the regret converges to a constant. Those simulation results are consistent with the theoretical upper bound in Theorem 5.

In order to show the benefit of the opportunistic exploration enabled by the availability of the immediate costs, we also investigate the performance of CC-UCB for the three cases under the assumption that the immediate costs are unknown beforehand. The corresponding cumulative regrets are plotted in Fig. 2(b). Compared with the results in Fig. 2(b), we note that the availability of the immediate cost information leads to orders of magnitude improvement of the regret for fixed T. The most dramatic contrast happens for case 3, where the regret grows sublinearly in T in Fig. 2(b), and it converges to a

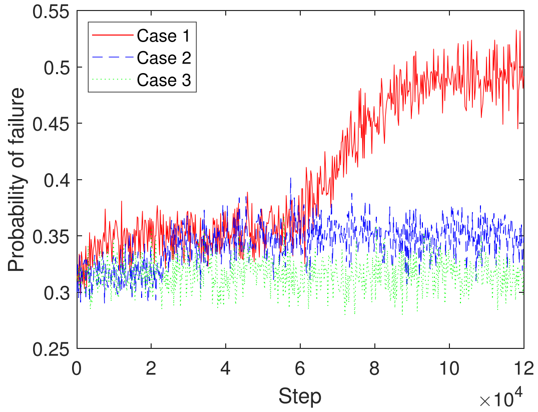


Fig. 3. Empirical probability of failure versus step.

constant in Fig. 2(a). We notice that for sufficiently large T, the regrets increase as $|\mathcal{S}_2|$ decreases. This is because with the given parameter selection, compared with arms with parameter setting c), those with parameter setting b) are harder to be distinguished from the good arms in I^*, leading to higher regret. This is also consistent with the theoretical bound in Theorem 2.

C. *Fail-Safe CCB*

Finally, we evaluate the performance of Fail-safe CCB through simulation. We consider a 3-arm bandits with expected reward $\boldsymbol{\theta} = [0.5, 0.3, 0.1]$. The cost of pulling each arm follows a Bernoulli distribution with expectation 0.35. Additionally, we assume that if the learner does not find any arm with state 1 after checking the list, an additional positive fixed cost would be incurred. We consider three possible cases, i.e., $C = 0, 1, 3$, respectively. According to Section V, under the optimal offline policy, $|I^*| = 1, 2, 3$, respectively.

We set $\epsilon = 10^{-5}$ and run Fail-safe CCB for 1000 sample paths. We track the empirical probability of failure in each step, and plot the curves in Fig. 3. As we note, for each of those three cases, the trend of the empirical probability of failure gradually increases as time increases, and eventually fluctuates around constants 0.5, 0.35 and 0.315, respectively. This is consistent with the fact that Fail-safe CCB performs exploration more aggressively at the beginning, thus resulting lower probability of failure; as time goes by, it gradually identifies the optimal offline policy, thus the probability of failure converges to the that under the optimal offline policy. Besides, as we expected, the probability of failure roughly decreases as C increases.

VII. Conclusion

In this paper, we studied a CCB model by taking the cost of pulling arms into the cascading bandits framework. We considered two different scenarios, depending on the availability of the immediate costs. For each scenario, we first explicitly characterized the optimal offline policy, and then developed a UCB based algorithm for the online setting. We analyzed the regret behavior of the proposed CC-UCB and CC-UCB2 algorithms, and showed that they are order-optimal. An extension to incorporate end-game failure was also studied. Experiments using both synthetic data and real-world data were carried out to evaluate the algorithms.

A. *Proof of Theorem 1*

Let $I = (I(1), I(2), \ldots, I(|I|))$ be the list that is presented by the player, where $|I| \leq K$. The expected net reward is

$$\mathbb{E}[r(I)] = \sum_{i=1}^{|I|} (\theta_{I(i)} - c_{I(i)}) \prod_{j=1}^{i-1} (1 - \theta_{I(j)}). \qquad (26)$$

If in the policy there exists $i \in I$ satisfying $\frac{\theta_{I(i+1)}}{c_{I(i+1)}} > \frac{\theta_{I(i)}}{c_{I(i)}}$ and $I(i), I(i+1) \in [K]$, then we can define another list I', which is only different from policy I by swapping the $I(i)$ and $I(i+1)$ position: $I' = (I(1), \ldots, I(i-1), I(i+1), I(i), I(i+2), \ldots, I(|I|))$, and $|I'| = |I|$. Then the difference between the expected net rewards of I and I' is

$$\begin{aligned}
\mathbb{E}[r(I')] - \mathbb{E}[r(I)] &= \left(\prod_{j=1}^{i-1} (1 - \theta_{I(j)})\right) \big[(\theta_{I(i+1)} - c_{I(i+1)}) \\
&\quad + (1 - \theta_{I(i+1)})(\theta_{I(i)} - c_{I(i)}) - (\theta_{I(i)} - c_{I(i)}) \\
&\quad - (1 - \theta_{I(i)})(\theta_{I(i+1)} - c_{I(i+1)})\big] \\
&= \left(\prod_{j=1}^{i-1} (1 - \theta_{I(j)})\right) (c_{I(i)}\theta_{I(i+1)} - c_{I(i+1)}\theta_{I(i)}) \\
&> 0,
\end{aligned} \qquad (27)$$

which implies that if the presented arms from $[K]$ are not in a descending order in $\frac{\theta_i}{c_i}$, then we can always create a new list that achieves better expected net reward by swapping positions of some arms.

Besides, the reward is the summation of $(\theta_{I(i)} - c_{I(i)}) \prod_{j=1}^{i-1} (1 - \theta_{I(j)})$. Then a term will be positive if $\theta_{I(i)} > c_{I(i)}$. As a result, the optimal offline policy must contain all $i : \theta_i > c_i$. Combining with (27), we reach the conclusion that the reward will be maximized by presenting the top L arms in a descending order based on $\frac{\theta_i}{c_i}$.

B. *Proof of Lemma 3*

Before we proceed to prove Lemma 3, we first introduce the following definitions.

Define a random variable $Z_{i,t}$ as follows:

$$Z_{i,t} = \begin{cases} 0, & \text{if } \mathbb{1}(\bar{\mathcal{E}}_t) = 0 \\ 0, & \text{if } \mathbb{1}(\bar{\mathcal{E}}_t) = 1, \text{ and } \exists j \in [K] \backslash \{i\}, X_{j,t} = 1 \\ 1, & \text{if } \mathbb{1}(\bar{\mathcal{E}}_t) = 1, \text{ and } \forall j \in [K] \backslash \{i\}, X_{j,t} = 0 \end{cases}. \qquad (28)$$

We can verify that

$$Z_{i,t} = \begin{cases} 0, & \text{if } \mathbb{1}(\bar{\mathcal{E}}_t) = 0 \\ \text{Bernoulli } (p_i), & \text{if } \mathbb{1}(\bar{\mathcal{E}}_t) = 1 \end{cases}, \qquad (29)$$

where $p_i := \frac{\prod_{j=1}^{K} (1 - \theta_j)}{(1 - \theta_i)}$.

As we will see later, we define $Z_{i,t}$ in such a way in order to lower bound the probability of the event "$\mathcal{E}_t$ is false and arm i is observed".

For any integer n, we define τ_n as the smallest step index such that $\sum_{t=1}^{\tau_n} \mathbb{1}(\bar{\mathcal{E}}_t) = n$. This definition implies that $\mathbb{1}(\bar{\mathcal{E}}_{\tau_n}) = 1$. Then, according to the definition of $Z_{i,t}$ in (29), Z_{i,τ_n} is Bernoulli (p_i). Since Z_{i,τ_n} is $\sigma(\{X_{k,\tau_n}\}_{k\in[K]})$ measurable and $\{X_{k,t}\}_{k\in[K]}$ are independent, $\{Z_{i,\tau_n}\}_n$ are independent. Therefore, $\{Z_{i,\tau_n}\}_{n=1}^{\infty}$ are i.i.d. Bernoulli random variables with parameter p_i.

We then denote $\Gamma_T := \sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)$, i.e., the total number of steps up to T when $\mathcal{E}_t$ is false. Then, we have the following observation.

Lemma 10: For all $i \in I^*$, $N_{i,T} \geq \sum_{t=1}^{T} Z_{i,t} = \sum_{n=1}^{\Gamma_T} Z_{i,\tau_n}$.

Proof: We note that

$$\begin{aligned}
N_{i,T} &= \sum_{t=1}^{T} \mathbb{1}(i \in \tilde{I}_t) && (30)\\
&\geq \sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(i \in \tilde{I}_t) && (31)\\
&\geq \sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(\forall j \in [K]\backslash\{i\}, X_{j,t} = 0\right) && (32)\\
&= \sum_{t=1}^{T} Z_{i,t}, && (33)
\end{aligned}$$

where (32) is based on the fact that arm i will be pulled only when the states of all arms in I_t listed before i are 0, and its probability is lower bounded by that of the extreme case when the states of all arms except i are 0; (33) follows from the definition of $Z_{i,t}$ in (28).

Since $\mathbb{1}(\bar{\mathcal{E}}_t) = \mathbb{1}(t \in \bigcup_{n=1}^{\infty}\{\tau_n\})$, we have $\sum_{t=1}^{T} Z_{i,t} = \sum_{n=1}^{\Gamma_T} Z_{i,\tau_n}$, and Lemma 10 follows. ■

Next, we are ready to prove Lemma 3.

Denote $\Delta_{i,j} := \frac{\left(\frac{\theta_i}{c_i} - \frac{\theta_j}{c_j}\right)c_j}{2\left(1+\frac{\theta_j}{c_j}\right)}$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\mathcal{B}_t)\right]\\
&\leq \sum_{j=2}^{L} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(\frac{U_{j^*,t}}{L_{j^*,t}} > \frac{\theta_{(j-1)^*}}{c_{(j-1)^*}}\right)\right]\\
&\leq \sum_{j=2}^{L} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(\frac{\theta_{j^*} + 2u_{j^*,t}}{c_{j^*} - 2u_{j^*,t}} > \frac{\theta_{(j-1)^*}}{c_{(j-1)^*}}\right)\right]\\
&= \sum_{j=2}^{L} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(N_{j^*,t} < \frac{4\left(1+\frac{\theta_{(j-1)^*}}{c_{(j-1)^*}}\right)^2 \alpha \log t}{\left(\frac{\theta_{(j-1)^*}}{c_{(j-1)^*}} - \frac{\theta_{j^*}}{c_{j^*}}\right)^2 c_{j^*}^2}\right)\right]\\
&= \sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(N_{j^*,\tau_n} < \frac{\alpha \log \tau_n}{\Delta^2_{(j-1)^*,j^*}}\right)\right]
\end{aligned}$$

$$= \sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(N_{j^*,\tau_n} < \frac{\alpha \log \tau_n}{\Delta^2_{(j-1)^*,j^*}}\right) \left(\mathbb{1}\left(\tau_n \leq 2n\right) + \mathbb{1}(\tau_n > 2n)\right)\right]. \tag{34}$$

We note that

$$\begin{aligned}
&\sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(N_{j^*,\tau_n} < \frac{\alpha \log \tau_n}{\Delta^2_{(j-1)^*,j^*}}\right)\mathbb{1}\left(\tau_n \leq 2n\right)\right]\\
&\leq \sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(N_{j^*,\tau_n} < \frac{\alpha \log (2n)}{\Delta^2_{(j-1)^*,j^*}}\right)\right] && (35)\\
&\leq \sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(\sum_{t=1}^{\tau_n} Z_{j^*,t} < \frac{\alpha \log (2n)}{\Delta^2_{(j-1)^*,j^*}}\right)\right] && (36)\\
&\leq \sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{1}\left(\sum_{t=1}^{n} Z_{j^*,\tau_t} < \frac{\alpha \log (2n)}{\Delta^2_{(j-1)^*,j^*}}\right)\right] && (37)\\
&= \sum_{j=2}^{L}\sum_{n=1}^{T} \mathbb{P}\left(\sum_{t=1}^{n} Z_{j^*,\tau_t} - np_{j^*} < \frac{\alpha \log (2n)}{\Delta^2_{(j-1)^*,j^*}} - np_{j^*}\right), && (38)
\end{aligned}$$

where (36) follows Lemma 10, (37) follows the fact that $\Gamma_T \leq T$.

When $\frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}} \geq e$, let

$$\zeta_j := \frac{4\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}} \log \frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}}.$$

Then, $\log(2\zeta_j) > 0$. Besides,

$$\begin{aligned}
\log(2\zeta_j) &= \log \frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}} + \log\log \frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}}\\
&\leq 2\log \frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{2\zeta_j}{\log(2\zeta_j)} &\geq \frac{\frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}} \log \frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}}}{2\log \frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}}}\\
&= \frac{4\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}},
\end{aligned}$$

which implies that for any $n \geq \zeta_j$, we must have

$$\frac{\alpha \log(2n)}{\Delta^2_{(j-1)^*,j^*}} \leq \frac{p_{j^*}}{2} n. \tag{39}$$

When $\frac{8\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}} < e$, we let $\zeta_j = 1$. Then,

$$\frac{2\zeta_j}{\log(2\zeta_j)} > e > \frac{4\alpha}{p_{j^*}\Delta^2_{(j-1)^*,j^*}},$$

which again implies (39) for any $n \geq \zeta_j$.

With the definition of ζ_j, we have

$$(38) \le \sum_{j=2}^{L} \left(\zeta_j + \sum_{n=\zeta_j+1}^{T} \mathbb{P}\left(\sum_{t=1}^{n} Z_{j^*,\tau_t} - np_{j^*} < -\frac{p_{j^*}}{2} n \right) \right)$$

$$\le \sum_{j=2}^{L} \left(\zeta_j + \sum_{n=\zeta_j+1}^{T} \exp\left(-2 \left(\frac{p_{j^*}}{2} \right)^2 n \right) \right) \tag{40}$$

$$\le \sum_{j=2}^{L} \left(\zeta_j + \frac{2}{p_{j^*}^2} \right), \tag{41}$$

where (40) follows from the fact that Z_{j^*,τ_t} are i.i.d. Bernoulli random variables with parameter p_{j^*} and Hoeffding's inequality.

Besides,

$$\sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(N_{j^*,\tau_n} < \frac{\alpha \log \tau_n}{\Delta^2_{(j-1)^*,j^*}} \right) \mathbb{1}\left(\tau_n > 2n \right) \right] \tag{42}$$

$$\le \sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(\tau_n > 2n \right) \right] \tag{43}$$

$$= \sum_{j=2}^{L} \mathbb{E}\left[\sum_{n=1}^{\Gamma_T} \mathbb{1}\left(\frac{\tau_n}{2} > \sum_{s=1}^{\tau_n} \mathbb{1}(\bar{\mathcal{E}}_t) \right) \right] \tag{44}$$

$$\le \sum_{j=2}^{L} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(\frac{t}{2} > \sum_{s=1}^{t} \mathbb{1}(\bar{\mathcal{E}}_t) \right) \right] \tag{45}$$

$$= \sum_{j=2}^{L} \sum_{t=1}^{T} \mathbb{P}\left[\sum_{s=1}^{t} \mathbb{1}(\mathcal{E}_s) > \frac{t}{2} \right], \tag{46}$$

where (44) is based on the definition of τ_n, (45) is due to the fact that $\Gamma_T \le T$, thus $\sum_{n=1}^{\Gamma_T} \mathbb{1}(\frac{\tau_n}{2} > \sum_{s=1}^{\tau_n} \mathbb{1}(\bar{\mathcal{E}}_t)) \le \sum_{t=1}^{T} \mathbb{1}(\frac{t}{2} > \sum_{s=1}^{t} \mathbb{1}(\bar{\mathcal{E}}_t))$.

We note that

$$\mathbb{E}[\mathbb{1}(\mathcal{E}_t)]$$

$$\le \sum_{k\in[K]} \left(\mathbb{P}\left[|\hat{\theta}_{k,t} - \theta_k| > u_{k,t} \right] + \mathbb{P}\left[|\hat{c}_{k,t} - c_k| > u_{k,t} \right] \right)$$

$$= \sum_{k\in[K]} \sum_{n=1}^{t} \left(\mathbb{P}\left[|\hat{\theta}_{k,t} - \theta_k| > \sqrt{\frac{\alpha \log t}{N_{k,t}}}, N_{k,t} = n \right] + \mathbb{P}\left[|\hat{c}_{k,t} - c_k| > \sqrt{\frac{\alpha \log t}{N_{k,t}}}, N_{k,t} = n \right] \right) \tag{47}$$

$$\le \sum_{k\in[K]} \sum_{n=1}^{t} 4 \exp\left(-2 \frac{\alpha \log t}{n} n \right) \tag{48}$$

$$= 4 \sum_{k\in[K]} t^{-2\alpha+1} = \frac{4K}{t^{2\alpha-1}} \le \frac{4K}{t^2}. \tag{49}$$

Therefore,

$$\mathbb{P}\left[\sum_{s=1}^{t} \mathbb{1}(\mathcal{E}_s) > \frac{t}{2} \right] \le \frac{\mathbb{E}\left[\left(\sum_{s=1}^{T} \mathbb{1}(\mathcal{E}_s) \right)^2 \right]}{(t/2)^2} \tag{50}$$

$$= \frac{4}{t^2} \left(\sum_{s=1}^{t} \mathbb{E}\left[\mathbb{1}(\mathcal{E}_s) \right] + 2 \sum_{1\le i<j\le t} \mathbb{E}\left[\mathbb{1}(\mathcal{E}_i) \mathbb{1}(\mathcal{E}_j) \right] \right) \tag{51}$$

$$\le \frac{4}{t^2} \left(\sum_{s=1}^{t} \mathbb{E}\left[\mathbb{1}(\mathcal{E}_s) \right] + 2 \sum_{1\le i<j\le t} \sqrt{\mathbb{E}\left[\mathbb{1}(\mathcal{E}_i)^2 \right] \mathbb{E}\left[\mathbb{1}(\mathcal{E}_j)^2 \right]} \right) \tag{52}$$

$$\le \frac{4}{t^2} \left(\sum_{s=1}^{t} \frac{4K}{s^2} + 2 \sum_{1\le i<j\le t} \frac{4K}{ij} \right) \tag{53}$$

$$= \frac{16K}{t^2} \left(\sum_{s=1}^{t} \frac{1}{s} \right)^2 < 16K \left(\frac{\log t + 1}{t} \right)^2, \tag{54}$$

where (50) follows from Chebyshev's inequality, (52) follows from Cauchy's inequality, and (53) follows from (49).

Plugging (54) into (46), we have

$$(42) \le (L-1) \sum_{t=1}^{\infty} 16K \left(\frac{\log t + 1}{t} \right)^2 < 16K \left(\frac{\pi^2}{6} + 1 + \log 2 + \frac{1}{3}(2 + \log^2 3 + 2\log 3) \right) := \xi_0. \tag{55}$$

Plugging (41) and (55) into (34), we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}(\mathcal{B}_t) \right] \le \sum_{j=2}^{L} \left(\zeta_j + \frac{1}{2p_{j^*}^2} \right) + \xi_0 := \zeta. \tag{56}$$

C. Proof of Lemma 6

Define a random variable $Z_{i,t}^m = Z_{i,t} \cdot \mathbb{1}(\mathbf{Y}_t \in \mathcal{Y}_m)$. Recall that $N_{i,t}^m$ is the number of times that arm i is pulled before step t under a cost vector in $\mathcal{Y}_m$. Keep the definitions of τ_n and Γ_T the same as in Appendix VII-B. Then, following similar argument as in the proof of Lemma 10, we can show that $N_{i,T}^m \ge \sum_{t=1}^{T} Z_{i,t}^m = \sum_{n=1}^{\Gamma_T} Z_{i,\tau_n}^m$ for all $i \in I^m$.

Focusing on the time slots when the cost vector $\mathbf{Y}_t \in \mathcal{Y}_m$, $\forall m$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}(\mathcal{B}_t) \mathbb{1}(\mathbf{Y}_t \in \mathcal{Y}_m) \right]$$

$$\le \sum_{j=2}^{|I^m|} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t) \mathbb{1}\left(\frac{U_{j^*,t}}{Y_{j^*,t}} > \frac{\theta_{(j-1)^*,t}}{Y_{(j-1)^*,t}} \right) \mathbb{1}(\mathbf{Y}_t \in \mathcal{Y}_m) \right]$$

$$\leq \rho_m \sum_{j=2}^{|I^m|} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(N_{j^*,t} < \frac{4\alpha \log t}{\Delta^2 l_{j^*}^2}\right)\right] \tag{57}$$

$$\leq \rho_m \sum_{j=2}^{|I^m|} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(N_{j^*,t}^m < \frac{4\alpha \log t}{\Delta^2 l_{j^*}^2}\right)\right], \tag{58}$$

where without ambiguity, we use j^* to denote the jth arm on the optimal list I^m.

Applying similar arguments as in Appendix VII-B on (58), we can show that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\mathcal{B}_t)\mathbb{1}(\mathbf{Y}_t \in \mathcal{Y}_m)\right]$$

$$\leq \rho_m \left(\sum_{j=2}^{|I^m|}\left(\zeta_j^m + \frac{2}{(\rho_m p_{j^*})^2}\right) + \xi_0\right) := \zeta^m,$$

where ζ_j^m equals $\frac{4\alpha}{\rho_m p_{j^*}\Delta^2 l_{j^*}^2} \log \frac{8\alpha}{\rho_m p_{j^*}\Delta^2 l_{j^*}^2}$ if $\frac{4\alpha}{\rho_m p_{j^*}\Delta^2 l_{j^*}^2} > e$, and it equals 1 otherwise.

D. Proof of Lemma 8

First, we denote $\mathcal{A}_t := \{\prod_{n=\lceil t/2 \rceil}^{t} \mathbb{1}(\bar{\mathcal{E}}_n) = 1\}$. Then, we have

$$\mathbb{P}\left[\hat{N}_{i,t}^m < \frac{\rho_m p_i t}{4}\right] \leq \mathbb{P}\left[\hat{N}_{i,t}^m < \frac{\rho_m p_i t}{4}, \mathcal{A}_t\right] + \mathbb{P}\left[\bar{\mathcal{A}}_t\right]. \tag{59}$$

When $t > 2K$,

$$\mathbb{P}\left[\bar{\mathcal{A}}_t\right] \leq \sum_{i=1}^{K} \sum_{s=\lceil \frac{t}{2} \rceil}^{t} \mathbb{P}\left[|\hat{\theta}_{i,s} - \theta_i| \geq u_{i,s}\right]$$

$$\leq \sum_{k=1}^{K} \sum_{s=\lceil \frac{t}{2} \rceil}^{t} \sum_{n=1}^{s} \mathbb{P}\left[|\hat{\theta}_{i,s} - \theta_i| \geq \sqrt{\frac{\alpha \log s}{N_{i,s}}}, N_{i,s} = n\right]$$

$$\leq K \sum_{s=\lceil \frac{t}{2} \rceil}^{t} \sum_{n=1}^{s} 2\exp(-2\frac{\alpha \log s}{n} n)$$

$$\leq 2K\left(\frac{t}{2}+1\right)\left(\frac{t}{2}\right)^{-2\alpha+1}. \tag{60}$$

Meanwhile, define $\hat{Z}_{i,t}^m$ as i.i.d. Bernoulli random variables with parameter $\rho_m p_i$ for $t \in [\lceil \frac{t}{2} \rceil, t]$. Then, when $\mathcal{A}_t$ is true, we can show that

$$\hat{N}_{i,t}^m \geq \sum_{n=\lceil \frac{t}{2} \rceil}^{t} \hat{Z}_{i,n}^m. \tag{61}$$

Therefore,

$$\mathbb{P}\left[\hat{N}_{i,t}^m < \frac{\rho_m p_i t}{4}, \mathcal{A}_t\right] \leq \mathbb{P}\left[\sum_{n=\lceil \frac{t}{2} \rceil}^{t} \hat{Z}_{i,n}^m < \frac{\rho_m p_i t}{4}, \mathcal{A}_t\right]$$

$$\leq \mathbb{P}\left[\sum_{n=\lceil \frac{t}{2} \rceil}^{t} \hat{Z}_{i,n}^m < \frac{\rho_m p_i t}{4}\right] \leq \exp\left(-\frac{\rho_m^2 p_i^2}{16} t\right). \tag{62}$$

Plugging (60) and (62) into (59), we have when $t > 2K$,

$$\mathbb{P}\left[\hat{N}_{i,t}^m < \frac{\rho_m p_i t}{4}\right]$$

$$\leq 2K\left(\frac{t}{2}+1\right)\left(\frac{t}{2}\right)^{-2\alpha+1} + \exp\left(-\frac{\rho_m^2 p_i^2}{16} t\right).$$

REFERENCES

[1] R. Zhou, C. Gan, J. Yan, and C. Shen, “Cost-aware cascading bandits,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3228–3234.

[2] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan, “Cascading bandits: Learning to rank in the cascade model,” in *Proc. 32nd Int. Conf. Mach. Learn*, 2015, pp. 767–776.

[3] J.-Y. Audibert and S. Bubeck, “Best arm identification in multi-armed bandits,” in *Proc. 23th Conf. Learn. Theory*, Haifa, Israel, Jun. 2010.

[4] S. Bubeck, R. Munos, and G. Stoltz, “Pure exploration in multi-armed bandits problems,” in *Proc. 20th Int. Conf. Algorithmic Learn. Theory*, 2009, pp. 23–37.

[5] S. Guha and K. Munagala, “Approximation algorithms for budgeted learning problems,” in *Proc. 39th Annu. ACM Symp. Theory Comput.*, New York, NY, USA, 2007, pp. 104–113.

[6] S. Shahrampour, M. Noshad, and V. Tarokh, “On sequential elimination algorithms for best-arm identification in multi-armed bandits,” *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4281–4292, Aug. 2017.

[7] C. Shen, “Universal best arm identification,” *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4464–4478, Sep. 2019.

[8] L. Tran-Thanh, A. C. Chapman, E. M. de Cote, A. Rogers, and N. R. Jennings, “Epsilon-first policies for budget-limited multi-armed bandits,” in *Proc. Assoc. Advancement Artif. Intell.*, 2010.

[9] L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings, “Knapsack based optimal policies for budget-limited multi-armed bandits,” in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1134–1140.

[10] A. Burnetas and O. Kanavetas, “Adaptive policies for sequential sampling under incomplete information and a cost constraint,” *Appl. Math. Informat. Mil. Sci.*, pp. 97–112, 2012.

[11] A. Burnetas, O. Kanavetas, and M. N. Katehakis, “Asymptotically optimal multi-armed bandit policies under a cost constraint,” *Probability Eng. Informational Sci.*, vol. 31, no. 3, pp. 284–310, 2017.

[12] W. Ding, T. Qiny, X.-D. Zhang, and T.-Y. Liu, “Multi-armed bandit with budget constraint and variable costs,” in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 232–238.

[13] Y. Xia, W. Ding, X.-D. Zhang, N. Yu, and T. Qin, “Budgeted bandit problems with continuous random costs,” in *Proc. Asian Conf. Mach. Learn.*, G. Holmes and T.-Y. Liu, Eds., vol. 45, Hong Kong, Nov. 2016, pp. 317–332.

[14] Y. Xia, H. Li, T. Qin, N. Yu, and T.-Y. Liu, “Thompson sampling for budgeted multi-armed bandits,” in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 3960–3966.

[15] Y. Xia, T. Qin, W. Ma, N. Yu, and T.-Y. Liu, “Budgeted multi-armed bandits with multiple plays,” in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 2210–2216.

[16] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, “Bandits with knapsacks,” in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, 2013, pp. 207–216.

[17] A. Sani, A. Lazaric, and R. Munos, “Risk-aversion in multi-armed bandits,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3275–3283.

[18] N. Galichet, M. Sebag, and O. Teytaud, “Exploration vs exploitation vs safety: Risk-aware multi-armed bandits,” in *Proc. 5th Asian Conf. Mach. Learn.*, Nov. 2013, pp. 245–260.

[19] S. Vakili and Q. Zhao, “Risk-averse multi-armed bandit problems under mean-variance measure,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 6, pp. 1093–1111, Sep. 2016.

[20] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: Iid rewards,” *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 968–976, Nov. 1987.

[21] J. Komiyama, J. Honda, and H. Nakagawa, “Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays,” in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 1152–1161.

[22] D. Kalathil, N. Nayyar, and R. Jain, “Decentralized learning for multiplayer multiarmed bandits,” *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2331–2345, Apr. 2014.

[23] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Combinatorial cascading bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1450–1458.

[24] S. Zong, H. Ni, K. Sung, N. R. Ke, Z. Wen, and B. Kveton, "Cascading bandits for large-scale recommendation problems," in *Proc. 32nd Conf. Uncertainty Artif. Intell.*, 2016, pp. 835–844.

[25] F. Radlinski, R. Kleinberg, and T. Joachims, "Learning diverse rankings with multi-armed bandits," in *Proc. 25th Intl. Conf. Mach. Learn.*, 2008, pp. 784–791.

[26] M. Streeter and D. Golovin, "An online algorithm for maximizing submodular functions," in *Proc. Adv. Neural Inf. Process. Syst. 21*, 2008, pp. 1577–1584.

[27] A. Slivkins, F. Radlinski, and S. Gollapudi, "Ranked bandits in metric spaces: Learning diverse rankings over large document collections," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 399–436, Feb. 2013.

[28] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, 2002.

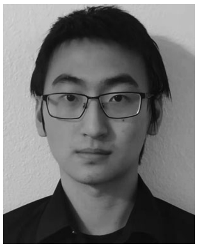
[29] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[30] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.

[31] H. Wu, X. Guo, and X. Liu, "Adaptive exploration-exploitation tradeoff for opportunistic bandits," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm Sweden, Jul. 2018, pp. 5306–5314.

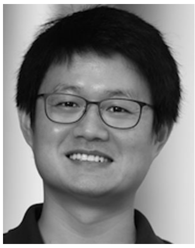
[32] "Internet mathematics," 2011. [Online]. Available: https://academy.yandex.ru/events/data_analysis/relpred2011/.

Chao Gan received the B.S. degree in statistics from the University of Science and Technology of China (USTC) in 2015. He is currently working towards the Ph.D. degree in electrical engineering at the Pennsylvania State University. His research interests include linear inverse problems, multi-armed bandits and their applications in communication and networking.



Ruida Zhou received the B.S. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Texas A&M University at College Station, College Station, TX, USA. His research interests include information theory and statistical learning.

Jing Yang (Member, IEEE) received the B.S. degree from the University of Science and Technology of China (USTC), and the M.S. and Ph.D. degrees from the University of Maryland, College Park, all in electrical engineering. She was a Postdoctoral Fellow at the University of Wisconsin-Madison, and an Assistant Professor in the Department of Electrical Engineering at the University of Arkansas. She is an Assistant Professor of electrical engineering at the Pennsylvania State University. Her research interests are in wireless communications and networking, statistical learning and signal processing, and information theory. Dr. Yang received an NSF CAREER award in 2015. She is now serving as an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.



Cong Shen (Senior Member, IEEE) received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, China in 2002 and 2004 respectively. He received the Ph.D. degree in electrical engineering from the University of California, Los Angeles, in 2009. From 2009 to 2014, He worked for Qualcomm Research in San Diego, CA, USA. From 2015 to 2019, he was with the School of Information Science and Technology, University of Science and Technology of China (USTC). He is currently an Assistant Professor at the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia. His general research interests are in the area of wireless communications and machine learning. He currently serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE WIRELESS COMMUNICATIONS LETTERS.