Cascading Bandits with Two-Level Feedback

Duo Cheng* Ruiquan Huang† Cong Shen‡ Jing Yang†
*Department of CS, Virginia Tech, Blacksburg, VA 24060
†School of EECS, The Pennsylvania State University, University Park, PA 16802
‡Department of ECE, University of Virginia, Charlottesville, VA 22904

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834892

***Abstract*—Motivated by the engineering application of efficient mobility management in ultra-dense wireless networks, we propose a novel cost-aware cascading bandit model with two-level actions. Compared with the standard cascading bandit model with a single-level action, this new model captures the real-world action sequence in mobility management, where the base station not only decides on an ordered neighbor cell list before measurement, but also executes the final handover decision to the target base station. We first analyze the optimal offline policy when the arm statistics are known beforehand. An online learning algorithm coined two-level Cost-aware Cascading UCB (CC-UCB) is then proposed to exploit the structure of the optimal offline policy with estimated arm statistics. Theoretical analysis shows that the cumulative regret under two-level CC-UCB scales logarithmically in time, which coincides with the asymptotic lower bound, thus is order-optimal. Simulation results corroborate the theoretical results and validate the effectiveness of two-level CC-UCB for mobility management.**

***Index Terms*—Cascading bandits, regret analysis, mobility management.**

I. Introduction

The learning goal of conventional cost-aware cascading bandit (CCB) [1] is the optimal *ordered* list of arms with unknown heterogeneous costs. This can be naturally applied to a mobility management problem where a neighbor cell list (NCL) is constructed by the serving base station (BS), and sent to the user equipment (UE) when the handover procedure is triggered [2], [3]. This design is particularly attractive in ultra-dense networks (UDN), where the possible target BSs are many while the construction of NCL is not delay-sensitive and can afford to take some time to converge [2], [4], [5].

Despite its natural fit to NCL construction, CCB does not take into account the handover decision and the corresponding outcome, which is the ultimate goal of mobility management in UDN. In this paper, we propose a new CCB model that has a two-level action structure, and the corresponding reward depends on the outcomes of the actions at both levels. This two-level CCB model is a better match to the UDN mobility management problem, where each time the UE is presented a ranked list of candidate BSs (first action). The UE will examine BSs one by one until one of the BSs has its status (e.g., reference signal received power (RSRP)) above a predefined threshold. The serving BS will send a handover request to this target BS upon receiving measurements from the UE (second action), and observe the handover outcome.

The first two authors contributed equally to this work. RH and JY were supported in part by NSF under grants 1956276, 2003131, and 2030026. CS was supported in part by NSF under grants 2002902 and 2029978.

Due to the two-level action structure, the optimal policy depends the statistics of the status of each BSs and the handover success rate. Correspondingly, the online learning algorithm needs to estimate both of those statistics. Since a BS may be examined multiple times before a handover request is triggered, the uncertainty levels in those two types of estimates are in general not balanced. How to take such unbalanced uncertainty levels into account to design an efficient online learning algorithm is a major challenge this work faces.

Our major contributions are summarized as follows. First, we formulate a novel cost-aware cascading bandits model with two-level actions. This model extends the classical cascading bandits model [6], [7], and may have direct applications in mobility managements in UDN and other areas. Second, we characterize the optimal *offline* policy when all statistics are known *a priori*. With the identified threshold structure of the optimal offline policy, we then design a two-level Cost-aware Cascading UCB (CC-UCB) algorithm to exploit this structure with empirically estimated statistical information, to solve the *online* learning problem when the statistics are unknown. We prove that the cumulative regret of the two-level CC-UCB algorithm scales in $O(\log T)$. Finally, we derive a matching asymptotic lower bound, proving that the two-level CC-UCB is order-optimal. Numerical results are reported to corroborate the theoretical analysis.

II. Two-Level CCB Model

We first describe the new two-level CCB model and highlight its correspondence in UDN mobility management. We consider a set of K arms (e.g., K BSs) denoted as $[K] = \{1, 2, \ldots, K\}$. Assume the status of each arm i at time t (e.g., RSRP of BS i), denoted by $X_{i,t}$, follows an unknown distribution ν_i. Each time, the learning agent (e.g., the serving BS) presents an ordered list of arms (e.g., NCL), denoted as $I_t := [I_t(1), \ldots, I_t(|I_t|)]$, to a player (e.g., UE). The player then starts examining the arms in I_t sequentially, until it finds the first arm whose status is above a predefined threshold γ (e.g., first BS that triggers the A4 event in 3GPP [8]). It would then stop examining the remaining arms in the list and send the obtained feedback to the agent. We define the probability to have $X_{i,t} \geq \gamma$ as p_i. Let $\tilde{I}_t \subseteq I_t$ be the list of arms that have been actually examined in step t, and $|\tilde{I}_t|$ be its size.

The aforementioned model captures the first-level action in cascading bandits. What is unique about the two-level CCB problem is that we also have access to a follow-up action: if $X_{I_t(|\tilde{I}_t|),t} \geq \gamma$, the agent will take another action on arm

$I_t(|\tilde{I}_t|)$ (e.g., send a handover request to the candidate BS $I_t(|\tilde{I}_t|)$) and observe its outcome (handover success $Y_{i,t}=1$ or failure $Y_{i,t}=0$ to the target BS i). In practice, even if the measured RSRP of the target BS is above the threshold, handover may still fail due to various reasons, e.g., measurement error or delay, BS load, etc. This second-level action and its corresponding feedback have not been utilized in prior designs [1], [2], which is the focus of this work. Model-wise, we assume whether the second action is successful is a *conditional* Bernoulli random variable. Specifically, given $X_{i,t} \geq \gamma$, the success probability of the second action is q_i. Otherwise, if $X_{i,t} < \gamma$, the success probability will always be zero.

Besides, we assume there is a cost associated with each arm pulling (e.g., energy consumption with each BS measurement). Denote $C_{i,t}$ as the cost of pulling arm i in step t. Without loss of generality, we assume $C_{i,t}$ is a bounded and non-negative independent and identically distributed (i.i.d.) random variable with $\mathbb{E}[C_{i,t}] = c_i$.

With a given ordered list I_t, $\tilde{I}_t$ is random and its realization depends on the observed $X_{i,t}$. Denote the *net reward* received by the learning agent at step t as

$$r_t := 1 - \prod_{i=1}^{|\tilde{I}_t|} \left(1 - Y_{I_t(i),t}\mathbb{1}(X_{I_t(i),t} \geq \gamma)\right) - \sum_{i=1}^{|\tilde{I}_t|} C_{I_t(i),t}.$$

Denote the observations up to step $t-1$ as $\mathcal{H}^{t-1}$. Then, without a priori statistics about $\{X_{i,t}\}$, $\{Y_{i,t}\}$ and $\{C_{i,t}\}$, our goal is to design an online algorithm to decide I_t based on observations obtained in previous steps $\mathcal{H}^{t-1}$, so as to minimize the *cumulative regret* $R(T) := Tr^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$, where r^* is maximum expected *net reward* if the statistics of $\{X_{i,t}\}_i$, $\{Y_{i,t}\}_i$ and $\{C_{i,t}\}_i$ were known beforehand. Besides, we also denote the per-step regret as $\mathsf{reg}_t := r^* - r_t$.

III. Optimal Offline Policy

Before we proceed to design the online learning algorithm and analyze its performance, we first identify the optimal offline policy assuming the statistics of arms and measuring costs are known a priori. For simplicity of the analysis, we make the following technical assumption:

Assumption 1 *$\frac{c_i}{p_i} \neq q_i$, for all $i \in [K]$.*

Our main result for the offline policy is given in the following theorem[1].

Theorem 1 *Arrange the arms in the decreasing order of $q_i - \frac{c_i}{p_i}$ and let L be the total number of arms with $q_i - \frac{c_i}{p_i} > 0$, i.e.,*

$$q_{1^*} - \frac{c_{1^*}}{p_{1^*}} \geq \ldots \geq q_{L^*} - \frac{c_{L^*}}{p_{L^*}} > 0 > q_{(L+1)^*} - \frac{c_{(L+1)^*}}{p_{(L+1)^*}} \geq \ldots \geq q_{K^*} - \frac{c_{K^*}}{p_{K^*}}.$$

[1]Due to page limit, we omit the proof of Theorem 1 in this paper. All missing proofs can be found in the supplementary material [9].

Algorithm 1 Two-level Cost-aware Cascading UCB

Input: α, γ.
Initialization: Examine all arms in $[K]$ once, and observe their states and costs.
while $t \leq T$ **do**
  for $i = 1 : K$ **do**
    $U_{i,t} = \hat{p}_{i,t} + u_{i,t}$;
    $V_{i,t} = \hat{q}_{i,t} + v_{i,t}$;
    $L_{i,t} = \max(\hat{c}_{i,t} - u_{i,t}, 0)$;
    if $\min(V_{i,t}, 1) - L_{i,t}/U_{i,t} > 0$ **then** $i \to I_t$;
    end if
  end for
  Rank arms in I_t in the descending order of $V_{i,t} - L_{i,t}/U_{i,t}$.
  for $i = 1 : |I_t|$ **do**
    Examine BS $I_t(i)$ and observe $X_{I_t(i),t}, C_{I_t(i),t}$;
    Add i to $\tilde{I}_t$;
    if $X_{I_t(i),t} \geq \gamma$ **then**
      Take second action on $I_t(i)$ and observe the outcome $Y_{I_t(i),t}$; break;
    end if
  end for
  Update $N_{i,t}$, $\hat{p}_{i,t}$, $\hat{c}_{i,t}$ for all $i \in \tilde{I}_t$;
  Update $M_{I_t(|\tilde{I}_t|),t}$ and $\hat{q}_{I_t(|\tilde{I}_t|)}$;
  $t = t + 1$;
end while

Then, I^* consists of the top L arms, and the corresponding optimal expected per-step net reward is $r^* = \sum_{i=1}^{L}(p_{i^*}q_{i^*} - c_{i^*})\prod_{j=1}^{i-1}(1-p_{j^*})$.

Compared with the optimal offline policy under the original CCB model in [1], a major difference is that the policy now depends on $q_i - \frac{c_i}{p_i}$ instead of $1 - \frac{c_i}{p_i}$, which captures the impact of two-level actions.

IV. Online Algorithm

A. Online Algorithm

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. The *two-level cost-aware cascading UCB* algorithm is described in Algorithm 1. We use $N_{i,t}$ to track the number of steps that arm i has been examined right before step t, and $\hat{p}_{i,t}$ to denote the sample average estimate of p_i. The UCB padding term 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. Besides, we use $M_{i,t}$ to track the number of steps that arm i has been chosen and attempted for handover right before step t, and $\hat{q}_{i,t}$ to denote the sample average estimate of q_i. We use $v_{i,t} := \sqrt{\frac{\alpha \log t}{M_{i,t}}}$ to denote the UCB padding term for q_i.

We point out that the main differences between the proposed two-level CC-UCB algorithm and the original CC-UCB in [1] include the estimation of the conditional success probability

q_i of the second action, the construction of its UCB, and the condition under which an arm should be included in I_t and subsequently undertake the second action. In particular, for the condition under which an arm should be included in I_t, we do not simply mimic the offline policy by checking whether $V_{i,t} - \frac{L_{i,t}}{U_{i,t}}$ is greater than 0. Rather, we first take the minimum between $V_{i,t}$ and 1, and then compare it with $\frac{L_{i,t}}{U_{i,t}}$. The reason we design the algorithm in this way is as follows. If we do not take the minimum between $V_{i,t}$ and 1, one extreme case would be that p_i is very close to 0 and q_i is very close to 1. Under the optimal offline policy, such an arm should not be included in I^*. Since p_i is very small, arm i would be rarely selected as a candidate arm. As a result, the corresponding padding term $v_{i,t}$ would be very large. Thus, the term $V_{i,t} - \frac{L_{i,t}}{U_{i,t}}$ would be very large as well, which implies that arm i would be a highly ranked arm in I_t. Such a decision deviates from the optimal offline policy, and incurs regret in almost every step. The adoption of the minimum operation prevents the algorithm from repetitively including such arms in I_t under the extreme scenario and improves its regret performance. On the other hand, we note that $V_{i,t}$ serves as an upper bound with high confidence on the true value of q_i, as q_i is a probability and must be upper bounded by one. Therefore, taking the minimum of $V_{i,t}$ and 1 provides a more reasonable upper bound for q_i.

B. Regret Upper Bound

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

Theorem 2 *The cumulative regret under Algorithm 1 is upper bounded as follows:*

$$R(T) \leq \sum_{i \in [K] \backslash I^*} (c_i + p_i(1-q_i))\, h_i \log T + O(1), \qquad (1)$$

where $[K]\backslash I^*$ includes all arms in $[K]$ except those in I^*, and h_i is a positive coefficient determined by p_i, q_i, c_i.

The remainder of this subsection is devoted to the proof sketch of Theorem 2. We present the three major steps (Parts I, II and III) and introduce several lemmas along the way.

At a high level, the proof is based on analyzing two error events: $\mathcal{E}_t$ and $\mathcal{B}_t$. The first happens when some parameter estimations are not in the corresponding confidence intervals at step t. The second event happens when the arms in I^* are not ordered correctly in I_t. We show that the probability of $\mathcal{E}_t$ and the probability of $\mathcal{B}_t \cap \bar{\mathcal{E}}_t$ are both negligible. The proof of Theorem 2 then completes after we show that the regret incurred under $\bar{\mathcal{E}}_t \cap \bar{\mathcal{B}}_t$ grows in $\log T$.

Part I: Analyzing $\mathcal{E}_t$. Mathematically, we define $\mathcal{E}_t := \{\exists i \in [K],\ |\hat{p}_{i,t} - p_i| > u_{i,t}$ or $|\hat{c}_{i,t} - c_i| > u_{i,t}$ or $|\hat{q}_{i,t} - q_i| > v_{i,t}\}$, i.e. there exists at least an arm whose sample average of status, second action success probability, or cost lies outside the corresponding confidence interval. Denote $\bar{\mathcal{E}}_t$ as the complement of $\mathcal{E}_t$. We first analyze the occurrence of $\mathcal{E}_t$.

Lemma 1 *Under Algorithm 1, we have* $\sum_{t=1}^{T} \mathbb{E}[\mathbb{1}(\mathcal{E}_t)] \leq \psi := K\left(1 + 2\pi^2\right)$.

Since ψ is a constant, indicating that $\bar{\mathcal{E}}_t$ happens linearly often, we can focus on the event $\bar{\mathcal{E}}_t$ in the remaining analysis. We establish that when $\bar{\mathcal{E}}_t$ happens, the candidate arms I_t defined in Algorithm 1 always contain I^*.

Lemma 2 *If $\mathbb{1}(\bar{\mathcal{E}}_t) = 1$, then, under Algorithm 1, all arms in I^* will be included in I_t.*

A few technical lemmas are needed to prove Theorem 2. The next lemma presents a lower bound of the probability p_i.

Lemma 3 *If $\mathbb{1}(\bar{\mathcal{E}}_t) = 1$, we have $p_i > c_i - 4u_{i,t}, \forall i \in I_t$.*

To ease the exposition, we introduce the following definitions:

$$\Delta_j := \frac{\left(q_{(j-1)^*} - \frac{c_{(j-1)^*}}{p_{(j-1)^*}}\right) - \left(q_{j^*} - \frac{c_{j^*}}{p_{j^*}}\right)}{1 + \frac{c_{j^*} + p_{j^*}}{p_{j^*}^2}}, \forall j \in [L]\backslash\{1\},$$

$$b := \min_{j \in [L]\backslash\{1\}} \sqrt{\frac{\Delta_{j^*}^2}{\Delta_{j^*}^2 + c_{j^*}}}, \qquad (2)$$

$$d := \min_{j \in [L]\backslash\{1\}} \sqrt{\frac{2p_{j^*}}{\left(\frac{1}{8p_{j^*}} + \frac{1}{\Delta_{j^*}^2}\right) \min_{i \in [L]\backslash\{1\}} c_{i^*}^2}}, \qquad (3)$$

$$\beta := \min(b, d).$$

For ease of argument, we assume $\Delta_j > 0$. The next lemma establishes a threshold for $N_{j,t}$ such that, when $N_{j,t}$ is above this threshold and hence sufficiently large, we can simultaneously guarantee (i) the probability that $M_{j,t}$ is small can be bounded using concentration inequalities; and (ii) the total regret summed over t steps converges. Both guarantees are critical in the proof of Lemma 7.

Lemma 4 *With the parameters defined above, when $N_{j^*,t} > \frac{16\alpha \log t}{\beta^2 \min_j c_{j^*}^2}$ and $\mathbb{1}(\bar{\mathcal{E}}_t) = 1$, the following inequalities hold:*

$$\frac{4\alpha \log t}{N_{j^*,t}\Delta_{j^*}^2} - p_{j^*} < 0,\ \forall j \in [L]\backslash\{1\}, \qquad (4)$$

$$\frac{32p_{j^*}^2}{\beta^2 \min_j c_{j^*}^2} - \frac{16p_{j^*}}{\Delta_{j^*}^2} > 2,\ \forall j \in [L]\backslash\{1\}. \qquad (5)$$

With the help of Lemma 4, we now present an upper bound for $M_{j,t}$ in Lemma 5.

Lemma 5 *If $n > \frac{16\alpha \log t}{\beta^2 \min_{j \in [L]\backslash\{1\}} c_{j^*}}$, for $j \in [L]\backslash\{1\}$, we have* $\mathbb{P}\left[M_{j^*,t} < \frac{4\alpha \log t}{\Delta_{j^*}^2}, N_{j^*,t} = n\right] \leq \exp\left(-2np_{j^*}^2 + \frac{16p_{j^*}\alpha \log t}{\Delta_{j^*}^2}\right)$.

The next lemma establishes a sublinear bound for the total number of pulls of a suboptimal arm when estimations are sufficiently accurate, which is useful in the regret analysis of Part III.

Lemma 6 *If $c_i > p_i q_i$, then there exists a positive coefficient h_i determined by p_i, q_i, c_i such that $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}\left(i \in \tilde{I}_t\right)\right] \leq h_i \log T$.*

Part II: Analyzing $\mathcal{B}_t$. Based on Lemma 2, when $\bar{\mathcal{E}}_t$ happens, we define $\mathcal{B}_t := \{\exists i^*, j^* \in I^*, i^* < j^*, \text{ s.t. } V_{i^*,t} - \frac{L_{i^*,t}}{U_{i^*,t}} < V_{j^*,t} - \frac{L_{i^*,t}}{U_{j^*,t}}\}$, 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, $\mathcal{B}_t$ should intuitively happen with small probability. This is characterized in Lemma 7.

Lemma 7 $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\mathcal{B}_t)\right] \leq \zeta = O(1)$, *where ζ is a constant depending on the problem parameters.*

The proof of Lemma 7 is based on the intuition that if $\bar{\mathcal{E}}_t$ is true, the estimations are close to the real values, and thus $\mathcal{B}_t$ should happen rarely. We note that Lemmas 3 to 5 are all used in the detailed proof of Lemma 7.

Lemma 8 *Consider an ordered list I_t that includes all arms from I^* with the same relative order as in I^*. Then, under Algorithm 1, $\mathbb{E}[\mathsf{reg}_t \mid \tilde{I}_t] \leq \sum_{i \in \tilde{I}_t \setminus I^*} (c_i + p_i(1-q_i))$.*

Part III: Putting Together. Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t)\mathbb{E}\left[\mathsf{reg}_t \mid \tilde{I}_t\right]\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t) \sum_{i \in \tilde{I}_t \setminus I^*} (c_i + p_i(1-q_i))\right] \\
&= \mathbb{E}\left[\sum_{i \in [K]\setminus I^*} \sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t)\mathbb{1}\left(i \in \tilde{I}_t\right)(c_i + p_i(1-q_i))\right] \\
&\leq \sum_{i \in [K]\setminus I^*} (c_i + p_i(1-q_i))\, h_i \log T
\end{aligned}
$$

where the last inequality comes from Lemma 6.

Denote δ^* as the largest possible per-step regret, which is bounded by $\sum_{i \in [K]} c_i$ and corresponds to the worst-case scenario that all arms are examined but the final reward is zero. Then, combining the results from Parts I and II, we have

$$
\begin{aligned}
R(T) &= \mathbb{E}\left[\sum_{t=1}^{T} [\mathbb{1}(\mathcal{E}_t) + \mathbb{1}(\bar{\mathcal{E}}_t)]\mathsf{reg}_t\right] \\
&\leq \delta^* \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(\mathcal{E}_t) + \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\mathcal{B}_t)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(\bar{\mathcal{E}}_t)\mathbb{1}(\bar{\mathcal{B}}_t)\mathbb{E}[\mathsf{reg}_t \mid \tilde{I}_t]\right] \\
&\leq \delta^*(\zeta + \psi) + \sum_{i \in [K]\setminus I^*} (c_i + p_i(1-q_i))\, h_i \log T,
\end{aligned}
$$

which completes the proof of Theorem 2.

C. Regret Lower Bound

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

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

Lemma 9 *For any ordered list I_t, the per-step regret in step t is lower bounded by $\mathbb{E}[\mathsf{reg}_t] \geq \mathbb{E}\left[\sum_{i \in \tilde{I}_t \setminus I^*} (c_i - p_i q_i)\right]$.*

Theorem 3 *Under any α-consistent policy,*

$$
\liminf_{T \to \infty} \frac{R(T)}{\log T} \geq \sum_{i \in [K]\setminus I^*} \max_{f_i'} \frac{c_i - p_i q_i}{D(f_i \parallel f_i')}, \tag{6}
$$

where f_i is the joint distribution of the state and success probability (i.e., X_i and Y_i) for arm i, and f_i' is a perturbed version of it such that $p_i' q_i' > c_i > p_i q_i$.

Remark: Consider a special case when $q_i = 1, \forall i$ (i.e., a handover attempt is always successful). In this case, f' should be designed such that $p_i < c < p_i'$. When X_i's are Bernoulli random variables at the same time, it degenerates to the case discussed in [1]. Thus, the lower bound in (6) recovers the lower bound in [1, Theorem 3].

Comparing Theorem 3 with Theorem 2, we conclude that Algorithm 1 achieves order-optimal regret performance.

V. Simulation Results

In this section we resort to numerical experiments to evaluate the performances of the proposed two-level CC-UCB algorithm. Besides, we also introduce two variants of the two-level CC-UCB algorithm in Algorithm 1 for comparison. First, we replace the condition $\min(V_{i,t}, 1) - L_{i,t}/U_{i,t} > 0$ in Line 8 of Algorithm 1 with $V_{i,t} - L_{i,t}/U_{i,t} > 0$. We denote this algorithm as Algorithm 2. Additionally, we replace the metric used to rank arms (i.e., $V_{i,t} - L_{i,t}/U_{i,t} > 0$) in Line 11 of Algorithm 1 with $\min(V_{i,t}, 1) - L_{i,t}/U_{i,t} > 0$, and denote the algorithm as Algorithm 3.

We evaluate the performances of these three algorithms in a 5-arm bandits setting with parameters $\mathbf{p} = [0.875, 0.75, 0.1, 0.5, 0.25]$ and $\mathbf{q} = [0.75, 0.7, 0.9, 0.55, 0.45]$. Besides, we assume the costs follow a uniform distribution $U(0.2, 0.4)$; hence $c = 0.3$ for all arms. These lead to $L = 2$ and $I^* = \{1, 2\}$ according to Theorem 1. Notice that the design of the third arm exactly follows the extreme case discussed in Section IV, with a low p_i and a relatively high q_i. We set $\alpha = 1.5$ and $\epsilon = 10^{-5}$, and run the algorithms for $T = 2 \times 10^5$ steps. The average of the cumulative regret over 20 runs is plotted in Fig. 1, where the error bar corresponds to one standard deviation of the regrets in 20 runs.

We have the following observations. First, all three algorithms achieve sublinear regret in the given setting. Algorithm

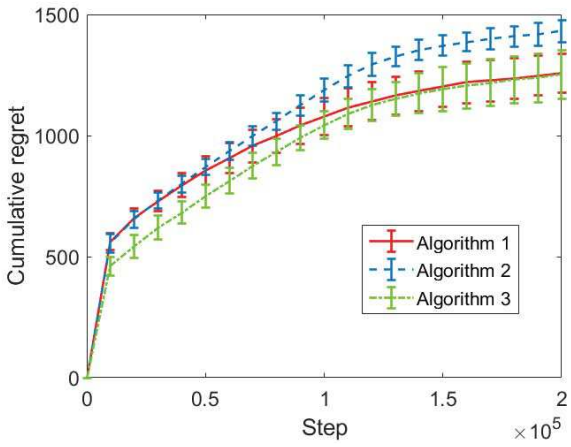

Fig. 1. Cumulative regret versus step

TABLE I
EXAMINATION DELAY VERSUS DIFFERENT c

Alg \ c	$c=0.25$		$c=0.3$		$c=0.35$		$c=0.4$	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Alg 1	1.1897	13	1.1830	13	1.1739	14	1.1666	36
Alg 2	1.1894	13	1.1825	21	1.1732	18	1.1667	24
Alg 3	1.1894	14	1.1825	23	1.1732	12	1.1662	24

1 achieves a significantly lower cumulative regret than Algorithm 2, especially when $t \geq 1 \times 10^5$, which corroborates our intuition that an extreme arm (e.g., the third arm) could incur much higher regret under algorithm 2. This validates the adoption of the $\min$ operator in Line 8 of Algorithm 1. Besides, we also note that the cumulative regrets under Algorithm 1 and Algorithm 3 are close to each other when t is sufficiently large. When t is small, the regret under Algorithm 3 is even lower than that under Algorithm 1. This indicates that taking the minimum between $V_{i,t}$ and 1 is also a good modification of the metric to rank the arms in I_t. We keep Algorithm 1 in its current form due to the simplicity of analyzing its theoretical performance, but in practice we can use a unified metric based on $\min(V_{i,t}, 1) - L_{i,t}/U_{i,t}$ to *add* and *rank* arms in the list.

We also study the influence of cost c on the average examination delay, which is captured by $|\tilde{I}_t|$, e.g,, how many BSs are examined during each handover process. The more arms are examined, the higher examination delay the system suffers from at this round. For that, we set the distribution of the random costs as $U(c-0.1, c+0.1)$ with different values of c shown in Table I. The rest of the setting stays the same. From Table I, we note that as c increases, the examination delay decreases slightly. This coincides with our intuition that with a larger value of c, I^* becomes shorter, leading to a lower examination delay on average.

Lastly, we evaluate the average success rate for varying c and report the results in Table II. We note that the success rate decreases as c increases, manifesting the trade-off between

TABLE II
SECOND ACTION SUCCESS RATE VERSUS DIFFERENT c

Alg \ c	$c=0.25$		$c=0.3$		$c=0.35$		$c=0.4$	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Alg 1	0.7333	13	0.7324	13	0.7317	13	0.7314	13
Alg 2	0.7330	12	0.7314	10	0.7309	9.5	0.7306	7.5
Alg 3	0.7327	9.2	0.7315	6.7	0.7308	7.2	0.7307	5.8

examination delay and success rate in this setting. Note that the standard deviation in Table I and Table II is in 10^{-4}.

VI. CONCLUSION

Motivated by the mobility management problem in ultra-dense networks, we have studied a new cost-aware cascading bandits problem with two-level actions. We explicitly identified the optimal offline policy, based on which an online learning algorithm termed two-level CC-UCB was proposed. We analyzed the performance of two-level UCB and showed that the cumulative regret scales in $O(\log T)$. A matching asymptotic lower bound was obtained as well, indicating two-level CC-UCB achieves order-optimal regret performance. Simulation results verified the sublinear regret of the proposed two-level CC-UCB algorithm. Although this work was motivated by the mobility management problem in UDN, the proposed two-level CC-UCB algorithm and its companion theoretical analyses are general and thus hold their own intellectual merit from the perspective of multi-armed bandits.

REFERENCES

[1] C. Gan, R. Zhou, J. Yang, and C. Shen, "Cost-aware cascading bandits," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3692–3706, 2020.
[2] C. Wang, R. Zhou, J. Yang, and C. Shen, "A cascading bandit approach to efficient mobility management in ultra-dense networks," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, (Invited Paper).
[3] Y. Li, E. Datta, J. Ding, N. Shroff, and X. Liu, "Bandit policies for reliable cellular network handovers in extreme mobility," *arXiv preprint arXiv:2010.15237*, 2020.
[4] C. Shen, C. Tekin, and M. van der Schaar, "A non-stochastic learning approach to energy efficient mobility management," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3854–3868, 2016.
[5] Y. Zhou, C. Shen, and M. van der Schaar, "A non-stationary online learning approach to mobility management," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1434–1446, Feb. 2019.
[6] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan, "Cascading bandits: Learning to rank in the cascade model," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 767–776.
[7] S. Zong, H. Ni, K. Sung, N. R. Ke, Z. Wen, and B. Kveton, "Cascading bandits for large-scale recommendation problems," *arXiv preprint arXiv:1603.05359*, 2016.
[8] 3GPP, "Telecommunication management; Configuration Management (CM); Notification Integration Reference Point (IRP); Requirements," TS 32.301, 2015.
[9] D. Cheng, R. Huang, J. Yang, and C. Shen, "Cascading bandits with two-level feedback," 2022. [Online]. Available: http://www.ee.psu.edu/yang/pub/ISIT22_supp.pdf
[10] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: I.I.D. rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.