Tight Gap-Dependent Regret Bounds and Problem-Independent Bounds for Cost-aware Cascading Bandits

Yuji Tamakoshi
The University of Tokyo and RIKEN AIP
uotstudent2001@g.ecc.u-tokyo.ac.jp

Shinji Ito
The University of Tokyo and RIKEN AIP
shinji@mist.i.u-tokyo.ac.jp

Abstract

We study cost-aware cascading bandits, where a learner selects an ordered subset of options, tests them sequentially until the first success, and pays the costs of all tested options. In this problem, regret comes from both testing inefficient options and placing options in a suboptimal order, but existing analyses do not separate these effects for inefficient options and therefore yield an inverse-square dependence on the gap $c_i - \theta_i$. We develop a regret decomposition based on intermediate policies that reorder the remaining suffix and remove inefficient options one position at a time. This allows us to quantify the incremental regret incurred when each option is tested. As a consequence, we show that the regret of CC-UCB admits a problem-dependent bound of $O(\sum_{i:\theta_i/c_i<1} \log T/(c_i - \theta_i))$ up to additive terms and a problem-independent bound of order $\tilde{O}(L\sqrt{T})$. We further prove that the minimax regret is bounded from below by $\Omega(\sqrt{LT})$ by reducing standard multi-armed bandits to a special case of the model. Finally, we propose CC-UCBv2, which removes the need to specify a positive lower bound on costs and handles zero-cost options by separating empirically zero-cost options from the others. Numerical experiments show the effect of misspecified cost lower bounds and demonstrate that the proposed modification can reduce regret in representative instances involving zero or misspecified costs.

1 Introduction

Cost-aware cascading bandits (CCB) constitute a class of problems that capture a common mathematical structure arising in applications such as communication systems and healthcare Zhou et al. [2018]. One motivating example is mobility management in wireless communication. In cellular handover under extreme mobility, neighboring cells are measured sequentially, and the serving cell faces a trade-off between taking more measurements and executing a timely handover Li et al. [2020]. Motivated by this structure, prior work models handover target selection as a sequential examination of candidate base stations from a neighbor cell list: a user equipment examines the recommended candidates in order and selects the first one that satisfies the handover condition Wang et al. [2019].

In this abstraction, each candidate base station corresponds to an option $i \in [L]$ in our model. Testing base station i at round t means checking whether it satisfies the connection or handover condition, represented by a Bernoulli outcome $X_{i,t} \in \{0, 1\}$ with success probability θ_i. This test also incurs a random time, signaling, or energy cost $Y_{i,t}$ with mean c_i. Thus, each option is characterized by its success probability θ_i and expected cost c_i, and its efficiency is $\rho_i := \theta_i/c_i$ when $c_i > 0$.

A related motivation comes from dynamic treatment allocation, where treatment effects can be modeled as Bernoulli outcomes and risks or burdens as costs Zhou et al. [2018]. The sequential

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

Table 1: Problem-dependent and problem-independent regret bounds for CCB with L options.

Paper	Bound	Problem Independent	Problem Dependent
Zhou et al. [2018]	Upper	none	$O(\sum_{i:\rho_i<1} c_i \frac{\log T}{(c_i-\theta_i)^2})^\dagger$
Zhou et al. [2018]	Lower	none	$\Omega\Big(\sum_{i:\rho_i<1} \frac{(c_i-\theta_i)\log T}{d(\theta_i;c_i)}\Big)^*$
Ours	Upper	$\tilde{O}(L\sqrt{T})$ Theorem 2	$O(\sum_{i:\rho_i<1} \frac{\log T}{c_i-\theta_i})^\dagger$ Theorem 1
Ours	Lower	$\Omega(\sqrt{LT})$ Theorem 3	none

† This upper bound ignores terms that do not depend on T.
* $d(\theta_i; c_i)$ is the KL divergence of Bernoulli distributions with means, θ_i and c_i.

testing-and-stopping structure in CCB further abstracts settings where candidate interventions are tried until a satisfactory outcome is achieved.

Abstracting these scenarios, at each round t, the learner selects an ordered subset $I_t = (I_t(1), \ldots, I_t(|I_t|))$ of options and tests them sequentially. The process terminates once a success is observed, so the subset of actually tested options, denoted by $\tilde{I}_t$, is random. The reward is defined as the success indicator minus the total incurred cost:

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

Hence, the learner must decide not only which options to test but also the order in which to test them.

From a theoretical perspective, CCB integrates two well-studied extensions of the multi-armed bandit (MAB) problem: bandits with costs Ding et al. [2013], Xia et al. [2015] and cascading feedback Kveton et al. [2015a,b]. It thereby provides a unified framework for settings where decisions incur costs and feedback is revealed sequentially until stopping. Given its generality, quantifying the intrinsic difficulty of this problem is of fundamental importance.

Despite its practical importance, theoretical understanding of CCB remains limited. One major difficulty lies in the fact that the ordering of options directly affects performance. We evaluate the performance of a learning algorithm by comparing its expected cumulative reward to that of an optimal offline policy, I^* that knows the true parameters in advance,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\big(r_t(I^*) - r_t(I_t)\big)\right]. \tag{2}$$

We call R_T the regret. CC-UCB Zhou et al. [2018] is based on the UCB algorithm for MAB problems. Specifically, it orders options according to the ratio between an upper confidence bound of the success probability θ_i and a lower confidence bound of the expected cost c_i. This promotes options with high efficiency ρ_i or high uncertainty to earlier positions, balancing exploration and exploitation. The regret in this setting arises from two sources: (i) selecting inefficient options that should not be tested, and (ii) misordering the options. Existing analyses cannot split these two sources for the case of inefficient options, which leads to a regret bound with an inverse-square dependence on the gap $c_i - \theta_i$.

In this work, we show that the regret of CC-UCB admits a problem-dependent bound of $O\Big(\sum_{i\in[L]:\,\theta_i/c_i<1} \frac{\log T}{c_i-\theta_i}\Big)$ (Theorem 1). This improves the bound of Zhou et al. [2018], which is of order $O(\sum_{i\in[L]:\theta_i/c_i<1} \frac{c_i \log T}{(c_i-\theta_i)^2})$. Moreover, our bound matches the regret lower bound of Zhou et al. [2018] in its dependence on the inefficiency gap $c_i - \theta_i$, up to constants and lower-order terms. Indeed, their regret lower bound contains $\Omega\Big(\sum_{i\in[L]:\theta_i/c_i<1} \frac{(c_i-\theta_i)\log T}{d(\theta_i;c_i)}\Big)$, where $d(\theta_i; c_i)$ is the Bernoulli KL divergence. In the small-gap regime, $d(\theta_i; c_i) = \Theta((c_i - \theta_i)^2)$, so this regret lower bound scales as $\Omega(\sum_{i\in[L]:\theta_i/c_i<1} \frac{\log T}{(c_i-\theta_i)})$. Furthermore, we show that the regret of CC-UCB admits a problem-independent bound of order $\tilde{O}(L\sqrt{T})$ (Theorem 2), and by considering the special case where all options always succeed, which reduces the problem to a standard MAB problem, we prove that the minimax regret is bounded from below by $\Omega(\sqrt{LT})$ (Theorem 3). These results are summarized in Table 1.

To obtain these results, we use an intermediate-policy decomposition that quantifies the regret caused by testing each option. At round t, the learner selects an ordered subset I_t. If the option at position i is less efficient than the one at position $i+1$, swapping them yields a policy I'_t whose expected reward improves by

$$\mathbb{E}_t[r_t(I'_t)] - \mathbb{E}_t[r_t(I_t)] = \prod_{j=1}^{i-1}(1-\theta_j)\big(\theta_{I_t(i+1)}c_{I_t(i)} - c_{I_t(i+1)}\theta_{I_t(i)}\big). \tag{3}$$

Directly comparing I_t with the optimal offline policy is challenging due to the cascading structure. Instead, we construct intermediate policies $I_t^{(i)}$ that agree with I_t up to position $i-1$, reorder the remaining suffix optimally, and remove inefficient options from that suffix. By sequentially comparing $I_t^{(i)}$ and $I_t^{(i+1)}$, we decompose the total regret into incremental differences incurred when each option is tested.

$$\begin{aligned}\mathbb{E}_t[(r(I_t^{(i)}) - r(I_t^{(i+1)}))] \le{}& \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}]c_{I_t(i)} \max_{k\in(\{I_t(j)|j>i\}\cap[L^*])\cup\{I_t(i)\}}\left(1-\frac{\rho_{I_t(i)}}{\rho_k}\right)\\ &+ \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)})\mathbf{1}\{I_t(i)\in[L]\setminus[L^*]\}.\end{aligned} \tag{4}$$

The first term measures the loss from testing $I_t(i)$ before a more efficient remaining optimal option; as $I_t(i)$ is sampled, this efficiency gap shrinks and the misordering regret diminishes. The decomposition also separates this loss from the regret of testing inefficient options.

We also propose a modified algorithm that does not require prior knowledge of a positive lower bound on the expected costs. CC-UCB assumes a known positive lower bound on the expected costs, which is often unavailable in practice. Our algorithm removes this assumption while preserving a problem-independent regret guarantee. Moreover, when distinguishing between zero-cost and nonzero-cost options, we show that simply checking whether a positive cost has ever been observed is more sample-efficient than relying only on confidence intervals. Numerical experiments demonstrate that this approach achieves smaller regret in representative instances involving zero or misspecified costs.

2 Related Literature

The MAB problem Robbins [1952] is a fundamental model for the trade-off between exploration and exploitation: a learner sequentially selects one of K arms with unknown reward distributions to maximize cumulative reward. Performance is typically evaluated by regret against an optimal policy that knows the true reward distributions. The UCB algorithm Auer et al. [2002] is a representative method whose regret admits a problem-dependent bound of $O(\sum_{i,\Delta_i>0}\frac{\log T}{\Delta_i})$ and a problem-independent bound of order $\tilde{O}(\sqrt{KT})$, where Δ_i is the gap between the reward expectation of arm i and that of the optimal arm. Lai and Robbins [1985] established an information-theoretic regret lower bound showing that, in the Bernoulli case, any uniformly good policy must incur regret $\Omega(\sum_{i:\Delta_i>0}\frac{\Delta_i\log T}{d(\mu_i;\mu^*)})$, where $\mu^* = \max_i \mu_i$ and $d(\mu_i;\mu^*)$ is the KL divergence between Bernoulli distributions with means μ_i and μ^*. For problem-independent regret, Auer et al. [1995] showed that the minimax regret is bounded from below by $\Omega(\sqrt{KT})$.

In many applications, pulling an arm incurs a cost, so the learner must maximize reward under resource constraints Ding et al. [2013], Xia et al. [2015]. Badanidiyuru et al. considered multiple types of costs and showed that, when the total budget is B, the regret admits an upper bound of $\tilde{O}(\sqrt{KB})$, while a matching lower bound of $\Omega(\sqrt{KB})$ also holds Badanidiyuru et al. [2013].

In bandits with multiple plays, the learner presents multiple options simultaneously Anantharam et al. [1987]; UCB-based algorithms have also been shown to be effective in this setting Chen et al. [2013]. Extensions that incorporate exploration costs into this setting have also been studied Xia et al. [2016].

Cascading bandits (CB) Kveton et al. [2015a,b] model sequential user behavior observed in applications such as online advertising and web search. For example, in web search, users examine presented items in order and stop once they find a satisfactory result. Unlike bandits with multiple plays, where feedback is observed for all presented items, CB provides only partial and stochastic feedback, and the presentation order directly affects which feedback is observed. In the standard CB setting, there

are L items with attraction probabilities $w_1, \dots, w_L$, and the learner recommends an ordered list of K items. After reindexing the items so that $w_1 \geq \cdots \geq w_L$, the optimal list consists of the top K items, and $\Delta_{e,K} := w_K - w_e$ denotes the gap between a suboptimal item $e > K$ and the K-th best item. UCB-based algorithms have also been proposed for this setting Kveton et al. [2015b], achieving regret of order $O(\sum_{e=K+1}^{L} \frac{\log T}{\Delta_{e,K}})$. For a hard instance with K optimal items of attraction probability p and $L - K$ suboptimal items of attraction probability $p - \Delta$, Kveton et al. [2015b] also established an information-theoretic regret lower bound $\Omega((L-K)\frac{\Delta(1-p)^{K-1}\log T}{d(p-\Delta;p)})$. Vial et al. [2022] proposed a Bernstein-type algorithm whose regret admits a bound of order $\tilde{O}(\sqrt{LT})$. They also showed that the minimax regret is bounded from below by $\Omega(\sqrt{LT})$ for the same setting.

Cost-aware cascading bandits (CCB) Zhou et al. [2018] extend this framework by incorporating both costs and cascading feedback. This model captures a general structure that appears in applications such as network routing and medical decision-making, and has also been studied from a more application-oriented perspective Cheng et al. [2022]. Although CC-UCB has been proposed for this problem, its theoretical analysis remains incomplete. In particular, the regret caused by misordering the options has not been sufficiently characterized. While analyses for CB can partially quantify the effect of incorrect ordering, these arguments cannot be directly extended to CCB. This is because, in standard CB, the expected reward depends only on the set of presented items and not on their order, although the order affects the observed feedback. In contrast, in CCB, the order influences not only the feedback but also the expected reward itself. Therefore, quantifying the reward difference induced by misordering is a central challenge in the theoretical analysis of this problem.

3 Problem Setting

In this work, we consider a stochastic bandit problem with L options. At each round t, when option $i \in [L]$ is selected, its outcome $X_{i,t} \in \{0, 1\}$ is drawn independently from a Bernoulli distribution with parameter θ_i, and its cost $Y_{i,t} \in [0, 1]$ is drawn independently from a distribution with mean c_i. When $c_i > 0$, we define the efficiency of option i as $\rho_i := \theta_i / c_i$. The analysis of the original CC-UCB algorithm is carried out under a positive lower bound on the expected costs, so all efficiencies are finite in that section. Zero-cost options are handled separately in Section 5.2, where we use an extended-real definition of efficiency. Without loss of generality, whenever efficiencies are used for an algorithmic or analytical statement, we reindex the option set in descending order of efficiency, so that

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_L, \tag{5}$$

with ties broken arbitrarily. Throughout the paper, option indices refer to this reindexed order.

At round t, the learning agent selects an ordered subset of options $I_t = (I_t(1), I_t(2), \dots, I_t(|I_t|))$ from the set $[L]$, and sequentially tests the options according to this order. The process terminates once a success is observed. Let $\tilde{I}_t$ denote the ordered subset of options that are actually tested at round t. For any $i \in [|\tilde{I}_t|]$, we have $\tilde{I}_t(i) = I_t(i)$.

The reward obtained at round t is defined as the success indicator minus the total incurred cost. That is,

$$r_t = 1 - \prod_{i=1}^{|\tilde{I}_t|} \big(1 - X_{\tilde{I}_t(i),t}\big) - \sum_{i=1}^{|\tilde{I}_t|} Y_{\tilde{I}_t(i),t}. \tag{6}$$

Equivalently, the reward can be written as

$$r_t = \sum_{k=1}^{|I_t|} \left(\prod_{j=1}^{k-1} \big(1 - X_{I_t(j),t}\big) \right) \Big(X_{I_t(k),t} - Y_{I_t(k),t} \Big). \tag{7}$$

We evaluate an algorithm by comparing its performance to that of always selecting the optimal ordered subset I^* that maximizes the expected reward. The performance gap is measured by the regret, defined as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \big(r_t(I^*) - r_t(I_t) \big) \right]. \tag{8}$$

From Theorem 1 of Zhou et al. [2018], the optimal policy I^* is given by the ordered subset consisting of all options with efficiency at least 1, arranged in descending order of efficiency. Let L^* denote the number of such options.

4 Analysis of CC-UCB

This section analyzes the original CC-UCB algorithm under the positive-cost assumption stated below. The zero-cost case is treated separately in Section 5.2. We define the following three events. Two of them are related to the sequential testing process, and the third concerns estimation errors of sample means.

Definition 1 (Reachability event)**.** *For $1 \le k \le |I_t|$, define*

$$\mathcal{R}_t(k) := \left\{ X_{I_t(1),t} = 0,\ X_{I_t(2),t} = 0,\ \ldots,\ X_{I_t(k-1),t} = 0 \right\}. \tag{9}$$

This event means that the learner reaches and tests the k-th option.

Definition 2 (Testing event)**.** *For each option i, define*

$$\mathcal{T}_{i,t} := \{ i \in I_t \} \cap \left\{ X_{j,t} = 0 \ \ \forall j \in I_t \text{ s.t. } I_t^{-1}(j) < I_t^{-1}(i) \right\}. \tag{10}$$

This event means that all options placed before i have failed, so option i is actually tested.

Definition 3 (Error event)**.** *At time t, define the error event $\mathcal{E}_t$ as*

$$\mathcal{E}_t := \left\{ \exists i \in [L],\ \left|\hat{\theta}_{i,t} - \theta_i\right| > \sqrt{\frac{1.5 \log t}{N_{i,t}}} \text{ or } \left|\hat{c}_{i,t} - c_i\right| > \sqrt{\frac{1.5 \log t}{N_{i,t}}} \right\}, \tag{11}$$

where $N_{i,t}$ is the number of observations of option i up to time t. On $\bar{\mathcal{E}}_t$, all estimates are sufficiently close to their true values.

4.1 Cost-Aware Cascading UCB (CC-UCB)

CC-UCB follows the standard UCB framework, balancing exploration and exploitation by optimistically estimating each option's efficiency.

4.1.1 Main Algorithmic Components

CC-UCB (Algorithm 1) maintains, for each option $i \in [L]$, the number of observations $N_{i,t}$, the empirical success probability $\hat{\theta}_{i,t}$, and the empirical expected cost $\hat{c}_{i,t}$. The algorithm first initializes these quantities by testing each option once. At each subsequent round t, it computes the confidence radius

$$u_{i,t} = \sqrt{\frac{1.5 \log t}{N_{i,t}}}. \tag{12}$$

Throughout the paper, we use the exploration constant 1.5, the smallest value allowed by the original analysis of Zhou et al. [2018] for radii of the form $\sqrt{\alpha \log t / N_{i,t}}$. The algorithm then constructs an upper confidence bound on the success probability and a lower confidence bound on the expected cost as

$$U_{i,t} = \hat{\theta}_{i,t} + u_{i,t}, \qquad L_{i,t} = \max\{\hat{c}_{i,t} - u_{i,t}, \epsilon\}. \tag{13}$$

The constant $\epsilon > 0$ ensures that the denominator is bounded away from zero. The algorithm then includes option i in the candidate set if its optimistic efficiency ratio satisfies

$$\frac{U_{i,t}}{L_{i,t}} \ge 1. \tag{14}$$

After constructing the candidate set, CC-UCB orders the selected options in descending order of $U_{i,t}/L_{i,t}$. The learner then tests the options sequentially according to this order and stops immediately once a success is observed. Only actually tested options are observed and updated using $(X_{i,t}, Y_{i,t})$; untested options remain unchanged. Thus, CC-UCB follows an optimistic principle: it favors options whose success probability may be large and whose cost may be small, while respecting the assumption that all costs are uniformly bounded below by ϵ.

4.1.2 Assumption and Theoretical Guarantees

For the input instance of CC-UCB, we make the following assumption:

Assumption 1. *There exists a known constant $\epsilon > 0$ such that $c_i > \epsilon$ for all options $i \in [L]$.*

Under this assumption, the following regret bounds hold.

Theorem 1. *Assume additionally that $\theta_i < 1$ for all options $i \in [L]$. The regret of CC-UCB satisfies*

$$R_T \leq \sum_{i \in [L], \frac{\theta_i}{c_i} < 1} \frac{72 \log T}{c_i - \theta_i} + O(1). \tag{15}$$

Here, the $O(1)$ term may depend on the problem instance but not on the horizon T.

This additional condition is used only in the problem-dependent analysis, where we invoke Lemma 3 of Zhou et al. [2018] to control the number of times optimal options are misordered. That lemma requires every option to be reached with positive probability when preceded by other options, which is ensured by $\theta_i < 1$ for all i. This condition is implicit in the corresponding argument of Zhou et al. [2018].

Theorem 2. *For sufficiently large T, the regret of CC-UCB satisfies*

$$R_T \leq \tilde{O}(L\sqrt{T}). \tag{16}$$

Proof sketch. The detailed proofs are provided in Appendix D. We summarize the main argument in three steps.

Good-event reduction. The proof first separates the regret on the confidence failure event $\mathcal{E}_t$ from the regret on the good event $\bar{\mathcal{E}}_t$. Standard UCB concentration bounds imply that the total contribution of $\mathcal{E}_t$ is only $O(1)$, so the main task is to control the regret under $\bar{\mathcal{E}}_t$.

Per-round decomposition. On the good event, the list I_t selected by CC-UCB contains every option in the optimal offline policy I^*. To compare CC-UCB's ordered list with I^*, we repeatedly sort suffixes and delete inefficient options. The adjacent-swap calculation in Appendix D yields

$$\begin{aligned} \mathbb{E}_t[r(I^*)] - \mathbb{E}_t[r(I_t)] \leq & \sum_{i=1}^{|I_t|-1} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] c_{I_t(i)} \max_{k \in (\{I_t(j) | j > i\} \cap [L^*]) \cup \{I_t(i)\}} \left(1 - \frac{\rho_{I_t(i)}}{\rho_k}\right) \\ & + \sum_{\substack{1 \leq i \leq |I_t| \\ I_t(i) \in [L] \setminus [L^*]}} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}). \end{aligned} \tag{17}$$

The first term is the loss from placing an option before a more efficient optimal option that appears later in the list. The second term is the loss from including inefficient options with $\rho_i < 1$.

Summing the per-round bound. The confidence intervals determine how often each term in the above decomposition can contribute to regret. Once option i has been sampled enough times, the optimistic ratio $U_{i,t}/L_{i,t}$ is accurate enough to prevent large efficiency mistakes involving i. For an inefficient option i, this gives at most $O(\log T/(c_i - \theta_i)^2)$ tests on the good event, hence $O(\log T/(c_i - \theta_i))$ regret after multiplying by the per-test loss $c_i - \theta_i$. The same threshold controls the regret from misordering inefficient options. Misorderings among optimal options contribute only instance-dependent constants, because sufficiently many observations separate each optimal option from the more efficient optimal options above it. Combining these contributions proves Theorem 1.

For Theorem 2, we use the same decomposition but truncate every gap-dependent count at T. Terms of the form $\min\{O(\log T/\Delta^2), T\}\Delta$ are at most $O(\sqrt{T \log T})$. Summing these gap-free contributions over the L options yields $\tilde{O}(L\sqrt{T})$.

Theorem 3. *For any algorithm, there exists a CCB instance such that*

$$R_T \geq \Omega(\sqrt{LT}). \tag{18}$$

Equivalently, the minimax regret is bounded from below by $\Omega(\sqrt{LT})$.

Proof. Consider the subclass of instances in which $\theta_i = 1$ for every option i and the costs are Bernoulli random variables. Since every tested option succeeds, each round stops after the first option in the chosen list. The remaining options in the list are neither tested nor observed. The learner's decision is therefore equivalent to choosing one option and receiving reward $1 - Y_{i,t}$. This is exactly a standard stochastic MAB problem with arm means $1 - c_i$. The minimax regret lower bound $\Omega(\sqrt{LT})$ for MAB problems [Auer et al., 1995] therefore applies to this subclass, and hence also to CCB. □

5 Extension of CC-UCB

5.1 Motivation for CC-UCBv2

CC-UCB assumes that the expected cost of every option is lower bounded by a known constant $\epsilon > 0$. This simplifies the ordering rule and analysis but can be restrictive: if the supplied lower bound exceeds an option's true cost, its efficiency may be underestimated, leading to suboptimal selection.

5.2 CC-UCBv2: Handling Zero and Near-Zero Costs

CC-UCBv2 (Algorithm 2) modifies the ordering rule of CC-UCB so that the algorithm remains well-defined even when some options may have zero expected cost. Options with $c_i = \theta_i = 0$ never produce a success and incur no cost, so they do not affect the reward and can be ignored in the analysis. For the remaining options, we use the following extended-real definition of efficiency:

$$\rho_i := \begin{cases} \dfrac{\theta_i}{c_i}, & \text{if } c_i > 0, \\ +\infty, & \text{if } c_i = 0 \wedge \theta_i > 0. \end{cases} \tag{19}$$

Ratios involving $+\infty$ are interpreted in the extended-real sense: if $\rho_i < +\infty$, then $\rho_i/(+\infty) = 0$.

As in CC-UCB, the algorithm maintains $N_{i,t}$, $\hat{\theta}_{i,t}$, and $\hat{c}_{i,t}$ for each option i, and initializes them by testing each option once. At round t, it computes

$$u_{i,t} = \sqrt{\frac{1.5 \log t}{N_{i,t}}}, \qquad U_{i,t} = \hat{\theta}_{i,t} + u_{i,t}, \qquad L_{i,t} = \max\{\hat{c}_{i,t} - u_{i,t}, 0\}. \tag{20}$$

Unlike CC-UCB, the lower confidence bound on the cost is not truncated by a positive constant. Therefore, the ratio $U_{i,t}/L_{i,t}$ may be undefined when $L_{i,t} = 0$. To handle this case, CC-UCBv2 includes option i in the candidate set if either

$$L_{i,t} = 0 \quad \text{and} \quad U_{i,t} > 0, \tag{21}$$

or

$$L_{i,t} > 0 \quad \text{and} \quad \frac{U_{i,t}}{L_{i,t}} \geq 1. \tag{22}$$

After selecting the candidate options, the theoretical version of CC-UCBv2 partitions them into two groups:

$$I_{L=0,t} = \{i \in I_t : L_{i,t} = 0\}, \tag{23}$$

and

$$I_{L>0,t} = \{i \in I_t : L_{i,t} > 0\}. \tag{24}$$

The first group, whose optimistic efficiency ratio is not well-defined, is ordered in descending order of $U_{i,t}$. The second group has $L_{i,t} > 0$ and is ordered in descending order of $U_{i,t}/L_{i,t}$. The final ordered list is obtained by concatenating these two groups:

$$I_t = I_{L=0,t} \| I_{L>0,t}. \tag{25}$$

The learner then tests the options in this order and stops upon the first success. This modification removes the need for prior knowledge of a positive cost lower bound while preserving the optimistic ordering principle whenever the efficiency ratio is well-defined.

A useful implementation refinement, CC-UCBv2-EZ, further splits $I_{L=0,t}$ by whether the empirical cost is still zero. It places options with $\hat{c}_{i,t} = 0$ before options with $\hat{c}_{i,t} > 0$. This refinement

is not needed for the regret guarantee below, but positive-cost options are expected to leave the empirical-zero group after only a small number of tests, whereas truly zero-cost options can remain there without increasing cost. To see this, fix an option with $c_i > 0$. While $\hat{c}_{i,t} = 0$, all cost observations of option i so far have been zero. For any cost distribution supported on $[0, 1]$ with mean c_i, we have $c_i = \mathbb{E}[Y_i] \leq \mathbb{P}(Y_i > 0)$; hence each additional test produces a nonzero cost with probability at least c_i and removes the option from the empirical-zero group. Thus the number of tests made while $\hat{c}_{i,t} = 0$ is stochastically dominated by a geometric waiting time with success probability c_i, and

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\mathbf{1}\{\hat{c}_{i,t} = 0\}\right] \leq \frac{1}{c_i}. \tag{26}$$

Thus positive-cost options leave the empirical-zero group quickly in expectation, whereas truly zero-cost options can remain there without incurring additional cost.

5.2.1 Theoretical Guarantees and Performance

The theorem below is stated for the conservative ordering rule of CC-UCBv2. This is the version for which we prove the regret guarantee. Its regret admits a problem-independent bound of order $\tilde{O}(L\sqrt{T})$. CC-UCBv2-EZ is an implementation refinement that keeps the same conservative ordering structure but separates empirically zero-cost options earlier in the list.

Theorem 4. *Let $L' := |\{i \in [L] : c_i = 0, \theta_i > 0\}|$ be the number of options with infinite efficiency. For sufficiently large T, the regret of the conservative CC-UCBv2 ordering rule satisfies*

$$R_T \leq \tilde{O}(L\sqrt{T}). \tag{27}$$

Proof. The detailed proof is given in Appendix E; we summarize the main modification. With the extended-real efficiency convention above, the adjacent-swap argument used for CC-UCB still gives the same per-round decomposition. If a zero-cost option with positive success probability appears after a positive-cost option, the efficiency-gap factor $1 - \rho_i/\rho_j$ is equal to one, so the corresponding swap loss is bounded by the coarse loss c_i. Lemma 1 controls how long such misorderings can occur: in the zero-cost case, the relevant threshold is $6 \log t/c_i^2$, which prevents a zero-cost option from remaining after a positive-cost option once $L_{i,t} > 0$.

Thus the proof of Theorem 2 applies to all finite-efficiency options. The infinite-efficiency options are optimal and have no more efficient option above them, so they do not contribute to either the misordering terms or the suboptimal-option term. Summing the same gap-free bounds gives the stated $\tilde{O}(L\sqrt{T})$ regret upper bound. □

6 Numerical Experiment

We conduct numerical experiments to compare CC-UCB, the conservative CC-UCBv2 ordering rule, and its empirical-zero refinement CC-UCBv2-EZ. Prior work on CCB has evaluated the effect of system parameters and conducted experiments using real-world click-log data [Zhou et al., 2018]. Our experiments instead focus on isolating phenomena that are directly tied to the theoretical and algorithmic questions studied in this paper: misspecified lower bounds on costs and zero-cost options. For each option i, the success probability and cost are assumed to follow Bernoulli distributions with parameters θ_i and c_i, respectively. We set $T = 200{,}000$ and report the average results over 100 independent simulations. Curves show mean cumulative regret, and error bars show one empirical standard deviation across runs. All experiments are lightweight CPU-based simulations reproducible on a standard laptop or desktop. Here, CC-UCBv2 denotes the conservative ordering rule analyzed in Theorem 4, whereas CC-UCBv2-EZ denotes the empirical-zero implementation refinement described in Section 5.2.

We first demonstrate that the proposed CC-UCBv2 variants outperform CC-UCB when the positive lower bound supplied to CC-UCB is misspecified. Consider the instance with $\boldsymbol{\theta} = \{0.9, 0.9, 0.9, 0.9, 0.9, 0.9\}$ and $\boldsymbol{c} = \{0.7, 0.6, 0.5, 0.4, 0.3, 0.2\}$, and set $\epsilon = 0.35$ for CC-UCB, which violates the assumed lower bound. The results are shown in Figure 1. The regret of CC-UCB grows linearly, since it fails to correctly distinguish between options 1 and 2.

Second, we show that CC-UCBv2-EZ can reduce regret in representative instances involving zero-cost options. Consider the instance $\boldsymbol{\theta} = \{0.5, 0.6, 0.7, 0.8, 0.9, 0.5\}$ and $\boldsymbol{c} = \{0.1, 0.1, 0.1, 0.1, 0.1, 0.0\}$. As shown in Figure 2, CC-UCBv2-EZ achieves lower regret than the original CC-UCBv2 algorithm.

We also report a supplementary experiment on the effect of specifying a positive lower bound on costs in Appendix A.

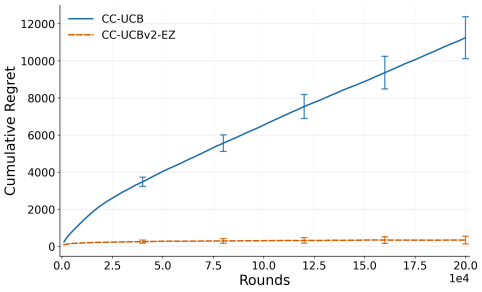


Figure 1: Algorithms under a misspecified cost lower bound.

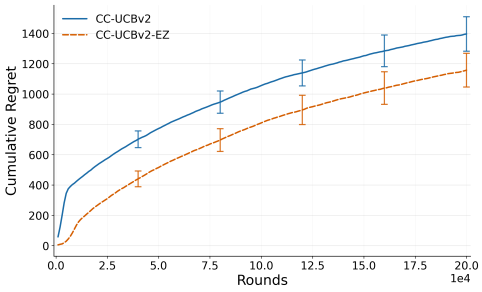


Figure 2: CC-UCBv2, and CC-UCBv2-EZ with a zero-cost option.

7 Limitations

Our results leave several limitations. First, there remains a gap between the problem-independent upper bound and the lower bound on the minimax regret. The current upper-bound analysis controls the maximum number of times each option is sampled. This ignores the fact that, under cascading feedback, options placed deeper in the list are tested less frequently because earlier successes stop the testing process. Exploiting this position-dependent observation structure may be necessary to tighten the bound.

Second, our upper-bound analysis is based on UCB-type confidence intervals. Sharper bounds may be possible with variance-adaptive confidence intervals, such as Bernstein-type bounds, as suggested by related work on cascading bandits [Vial et al., 2022]. However, in CCB the learner must estimate both success probabilities and costs, and these two quantities enter the ordering rule through a ratio. This makes a direct Bernstein-style extension technically nontrivial.

Third, the numerical experiments are intended to illustrate representative phenomena, such as misspecified lower bounds on costs and the behavior of zero-cost options, rather than to provide an exhaustive empirical comparison.

Finally, the broader impact of this work depends on the application in which sequential testing is deployed. In high-stakes domains, deployment should validate the modeling assumptions and cost definitions, since misuse could amplify decision errors.

8 Conclusion

In this paper, we develop a regret decomposition for the CCB problem that separates the regret caused by testing inefficient options from the regret caused by misordering options. This yields a problem-dependent bound on the regret with an improved dependence on the inefficiency gap $c_i - \theta_i$, as well as a problem-independent bound for CC-UCB. In addition, by considering instances consisting only of options with $\theta_i = 1$, for which the problem reduces to a standard MAB setting, we show that the minimax regret is bounded from below.

We further remove the assumption $\epsilon > 0$ on the expected cost and distinguish zero-cost from nonzero-cost options by detecting whether the empirical cost has ever been observed to be nonzero. We validate the effectiveness of this approach through numerical experiments.

References

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 Trans. Automat. Contr.*, 32(11):

968–976, November 1987.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1995. doi: 10.1109/SFCS.1995.492488.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, page 207–216. IEEE, October 2013. doi: 10.1109/focs.2013.30. URL `http://dx.doi.org/10.1109/FOCS.2013.30`.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. *ICML*, 28(1):151–159, February 2013.

Duo Cheng, Ruiquan Huang, Cong Shen, and Jing Yang. Cascading bandits with two-level feedback. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1892–1896, 2022. doi: 10.1109/ISIT50566.2022.9834892.

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. *Proc. Conf. AAAI Artif. Intell.*, 27(1):232–238, June 2013.

B Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *Neural Inf Process Syst*, abs/1507.04208, July 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 420–429, Arlington, Virginia, USA, 2014. AUAI Press. ISBN 9780974903910.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 767–776, Lille, France, 07–09 Jul 2015b. PMLR. URL `https://proceedings.mlr.press/v37/kveton15.html`.

T L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6 (1):4–22, March 1985.

Yuanjie Li, Esha Datta, Jiaxin Ding, Ness Shroff, and Xin Liu. Bandit policies for reliable cellular network handovers in extreme mobility. *arXiv [cs.LG]*, October 2020.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. New Ser. Am. Math. Soc.*, 58(5):527–535, 1952.

Daniel Vial, S Sanghavi, S Shakkottai, and R Srikant. Minimax regret for cascading bandits. *Neural Inf Process Syst*, abs/2203.12577:29126–29138, March 2022.

Chao Wang, Ruida Zhou, Jing Yang, and Cong Shen. A cascading bandit approach to efficient mobility management in ultra-dense networks. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, October 2019.

Yingce Xia, Wenkui Ding, Xudong Zhang, Nenghai Yu, and Tao Qin. Budgeted bandit problems with continuous random costs. *Asian Conf Mach Learn*, 45:317–332, 2015.

Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits with multiple plays. *Int Jt Conf Artif Intell*, pages 2210–2216, July 2016.

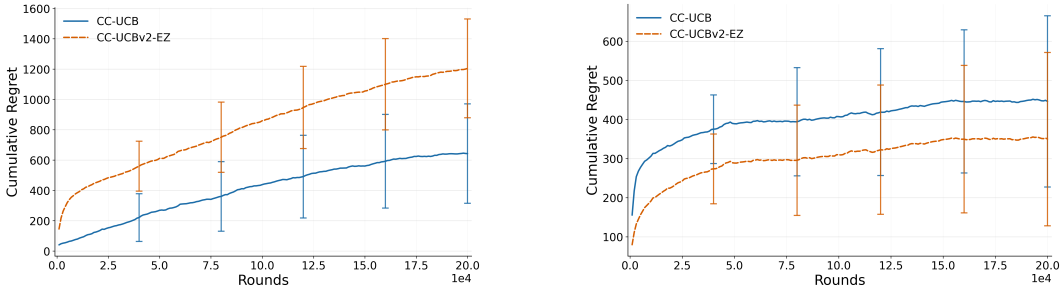
Ruida Zhou, Chao Gan, Jing Yang, and Cong Shen. Cost-aware cascading bandits. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3228–3234. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/448. URL `https://doi.org/10.24963/ijcai.2018/448`.

Acknowledgments and Disclosure of Funding

SI is supported by JSPS KAKENHI Grant Number JP25K03184 and by JST PRESTO, Japan, Grant Number JPMJPR2511.

A Additional Numerical Experiment

Zhou et al. [2018] experimentally showed that incorporating a positive lower bound on expected costs can improve regret. Here we include two representative instances illustrating that the empirical effect of this lower bound can depend on the instance. We consider $\boldsymbol{\theta} = \{0.8, 0.7, 0.6, 0.5, 0.4, 0.3\}$, $\boldsymbol{c} = \{0.55, 0.55, 0.55, 0.55, 0.55, 0.55\}$ with $\epsilon = 0.5$, and $\boldsymbol{\theta} = \{0.9, 0.9, 0.9, 0.9, 0.9, 0.9\}$, $\boldsymbol{c} = \{0.7, 0.6, 0.5, 0.4, 0.3, 0.2\}$ with $\epsilon = 0.19$. In the former case, Figure 3a shows that specifying ϵ reduces regret, whereas in the latter case, Figure 3b shows that it increases regret. These results are intended only as an empirical observation; we do not provide a theoretical characterization of when specifying such a lower bound is beneficial.



(a) Specifying ϵ reduces the regret of CC-UCB.

(b) Specifying ϵ increases the regret of CC-UCB.

Figure 3: Supplementary experiment on specifying a positive cost lower bound.

B Algorithms

This section presents the detailed descriptions of the algorithms analyzed in this paper: CC-UCB and CC-UCBv2. Both algorithms follow the upper confidence bound (UCB) paradigm and adaptively select and order options to balance exploration and exploitation in the cost-aware cascading bandits setting. In both algorithms, the confidence radius uses the fixed constant 1.5, the smallest value allowed by the theoretical analysis of Zhou et al. [2018]. In the usual notation $u_{i,t} = \sqrt{\alpha \log t / N_{i,t}}$, this corresponds to setting the exploration constant to $\alpha = 1.5$.

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

Require: $\epsilon > 0$

Initialization: test each option $i \in [L]$ once; observe $(X_{i,1}, Y_{i,1})$ and set
    $N_{i,1} \leftarrow 1,\ \hat{\theta}_{i,1} \leftarrow X_{i,1},\ \hat{c}_{i,1} \leftarrow Y_{i,1}$.
for $t = 2, 3, \ldots$ **do**
    $I_t \leftarrow \emptyset$
    for $i = 1, 2, \ldots, L$ **do**
        $u_{i,t} \leftarrow \sqrt{\dfrac{1.5 \log t}{N_{i,t}}}$ ▷ Confidence radius
        $U_{i,t} \leftarrow \hat{\theta}_{i,t} + u_{i,t}$ ▷ Upper confidence bound on θ_i
        $L_{i,t} \leftarrow \max\{\hat{c}_{i,t} - u_{i,t},\ \epsilon\}$ ▷ Lower confidence bound on c_i
        if $\dfrac{U_{i,t}}{L_{i,t}} \geq 1$ **then**
            $I_t \leftarrow I_t \cup \{i\}$ ▷ Include options with $\frac{U_{i,t}}{L_{i,t}} \geq 1$
        end if
    end for
    Sort options in I_t in descending order of $\dfrac{U_{i,t}}{L_{i,t}}$ ▷ Efficiency-based ordering
    $\tilde{I}_t \leftarrow \emptyset$
    for $k = 1, 2, \ldots, |I_t|$ **do**
        test option $I_t(k)$; observe $(X_{I_t(k),t}, Y_{I_t(k),t})$
        $\tilde{I}_t \leftarrow \tilde{I}_t \cup \{I_t(k)\}$
        if $X_{I_t(k),t} = 1$ **then**
            break ▷ Stop on first success
        end if
    end for
    for all $i \in \tilde{I}_t$ **do**
        $N_{i,t+1} \leftarrow N_{i,t} + 1$
        $\hat{\theta}_{i,t+1} \leftarrow \dfrac{N_{i,t}\hat{\theta}_{i,t} + X_{i,t}}{N_{i,t+1}}$ ▷ Update success probability estimate
        $\hat{c}_{i,t+1} \leftarrow \dfrac{N_{i,t}\hat{c}_{i,t} + Y_{i,t}}{N_{i,t+1}}$ ▷ Update cost estimate
    end for
    for all $i \notin \tilde{I}_t$ **do**
        $N_{i,t+1} \leftarrow N_{i,t},\ \hat{\theta}_{i,t+1} \leftarrow \hat{\theta}_{i,t},\ \hat{c}_{i,t+1} \leftarrow \hat{c}_{i,t}$ ▷ No update for untested options
    end for
end for

Algorithm 2 CC-UCBv2 — Conservative Ordering (Well-Defined for Zero-Cost Options)

Initialization: test each option $i \in [L]$ once; observe $(X_{i,1}, Y_{i,1})$ and set
    $N_{i,1} \leftarrow 1,\ \hat{\theta}_{i,1} \leftarrow X_{i,1},\ \hat{c}_{i,1} \leftarrow Y_{i,1}$.
for $t = 2, 3, \ldots$ **do**
    $I_t \leftarrow \emptyset$
    for $i = 1, 2, \ldots, L$ **do**
        $u_{i,t} \leftarrow \sqrt{\dfrac{1.5 \log t}{N_{i,t}}}$
        $U_{i,t} \leftarrow \hat{\theta}_{i,t} + u_{i,t}$
        $L_{i,t} \leftarrow \max\{\hat{c}_{i,t} - u_{i,t},\ 0\}$ ▷ No lower bound on cost
        if $\big(L_{i,t} = 0 \ \wedge\ U_{i,t} > 0\big) \ \vee\ \big(L_{i,t} > 0 \ \wedge\ \frac{U_{i,t}}{L_{i,t}} \geq 1\big)$ **then**
            $I_t \leftarrow I_t \cup \{i\}$ ▷ More flexible inclusion criterion
        end if
    end for
    (Conservative Ordering:) Partition I_t into two disjoint sets:
        $I_{L=0} \leftarrow \{i \in I_t : L_{i,t} = 0\}$ ▷ Undefined ratio group
        $I_{L>0} \leftarrow \{i \in I_t : L_{i,t} > 0\}$ ▷ Well-defined ratio group
    Sort options in $I_{L=0}$ in descending order of $U_{i,t}$ ▷ By success probability
    Sort options in $I_{L>0}$ in descending order of $\dfrac{U_{i,t}}{L_{i,t}}$ ▷ By efficiency ratio
    Re-define I_t as the concatenation:
        $I_t \leftarrow [I_{L=0}] \,\|\, [I_{L>0}]$ ▷ Hierarchical ordering
    $\tilde{I}_t \leftarrow \emptyset$
    for $k = 1, 2, \ldots, |I_t|$ **do**
        test option $I_t(k)$; observe $(X_{I_t(k),t}, Y_{I_t(k),t})$
        $\tilde{I}_t \leftarrow \tilde{I}_t \cup \{I_t(k)\}$
        if $X_{I_t(k),t} = 1$ **then**
            break
        end if
    end for
    for all $i \in \tilde{I}_t$ **do**
        $N_{i,t+1} \leftarrow N_{i,t} + 1$
        $\hat{\theta}_{i,t+1} \leftarrow \dfrac{N_{i,t}\hat{\theta}_{i,t} + X_{i,t}}{N_{i,t+1}}$
        $\hat{c}_{i,t+1} \leftarrow \dfrac{N_{i,t}\hat{c}_{i,t} + Y_{i,t}}{N_{i,t+1}}$
    end for
    for all $i \notin \tilde{I}_t$ **do**
        $N_{i,t+1} \leftarrow N_{i,t},\ \hat{\theta}_{i,t+1} \leftarrow \hat{\theta}_{i,t},\ \hat{c}_{i,t+1} \leftarrow \hat{c}_{i,t}$
    end for
end for

C Auxiliary Lemmas

Lemma 1 (Threshold for misordering)**.** *For option i and $j \in [(i-1) \wedge L^*]$, define $n_{j,i,t}$ by*

$$n_{j,i,t} := \begin{cases} +\infty & \text{if } \rho_j = \rho_i, \\ \frac{6\log t}{c_i^2} & \text{if } c_j = 0 \text{ and } \rho_j \neq \rho_i, \\ \frac{6(1+\rho_j)^2 \log t}{(\rho_j - \rho_i)^2 c_i^2} & \text{otherwise}. \end{cases} \tag{28}$$

For any $j \in [(i-1) \wedge L^*]$, if the event $\bar{\mathcal{E}}_t$ holds at time t and $n_{j-1,i,t} < N_{i,t} \leq n_{j,i,t}$, then no option more efficient than j can appear after i under the ordering rule of CC-UCB, or under CC-UCBv2 when zero-cost options are allowed.

Proof. It suffices to consider options j that are more efficient than i. First suppose that $c_j > 0$. Since the event $\bar{\mathcal{E}}_t$ holds, if option j is placed after i, then

$$\mathbf{1}\left\{\frac{U_{i,t}}{L_{i,t}} \geq \frac{U_{j,t}}{L_{j,t}}\right\} \leq \mathbf{1}\left\{\frac{U_{i,t}}{L_{i,t}} \geq \frac{\theta_j}{c_j}\right\} \tag{29}$$

$$\leq \mathbf{1}\left\{c_i - 2u_{i,t} \leq 0 \ \cup \ \frac{\theta_i + 2u_{i,t}}{c_i - 2u_{i,t}} \geq \frac{\theta_j}{c_j}\right\} \tag{30}$$

$$= \mathbf{1}\left\{N_{i,t} \leq \max\left\{\frac{6\log t}{c_i^2}, \frac{6(1+\rho_j)^2 \log t}{(\rho_j - \rho_i)^2 c_i^2}\right\}\right\} \tag{31}$$

$$= \mathbf{1}\left\{N_{i,t} \leq \frac{6(1+\rho_j)^2 \log t}{(\rho_j - \rho_i)^2 c_i^2}\right\}. \tag{32}$$

Thus the desired threshold holds when $c_j > 0$.

Now suppose that option j has zero expected cost. Under the CC-UCBv2 ordering rule, such an option is placed before any positive-cost option whose lower confidence bound on the cost is positive. Therefore, if j is placed after i, then $L_{i,t} = 0$. On $\bar{\mathcal{E}}_t$,

$$\mathbf{1}\{L_{i,t} = 0\} \leq \mathbf{1}\{c_i - 2u_{i,t} \leq 0\} \tag{33}$$

$$= \mathbf{1}\left\{N_{i,t} \leq \frac{6\log t}{c_i^2}\right\}. \tag{34}$$

Hence a zero-cost option j can appear after i only when $N_{i,t} \leq 6\log t / c_i^2 = n_{j,i,t}$.

Therefore, if $n_{j-1,i,t} < N_{i,t} \leq n_{j,i,t}$, then no option more efficient than j can appear after i. □

Lemma 2 (Selection of a suboptimal option)**.** *If the event $\bar{\mathcal{E}}_t$ holds at time t, then for any suboptimal option i, if*

$$N_{i,t} > \frac{24\log t}{(c_i - \theta_i)^2}, \tag{35}$$

it follows that $i \notin I_t$.

Proof. The event that a suboptimal option i is included in I_t satisfies

$$\mathbf{1}\{U_{i,t} \geq L_{i,t}\} \leq \mathbf{1}\{\theta_i + 2u_{i,t} \geq c_i - 2u_{i,t}\}$$

$$\leq \mathbf{1}\left\{N_{i,t} \leq \frac{24\log t}{(c_i - \theta_i)^2}\right\}. \tag{36}$$

Thus, this event is contained in an event depending only on the number of observations $N_{i,t}$ up to time t. Therefore, if $\bar{\mathcal{E}}_t$ holds and the number of times option i has been observed exceeds $\frac{24\log t}{(c_i-\theta_i)^2}$, then we have $U_{i,t} < L_{i,t}$, which implies $i \notin I_t$. □

Lemma 3 (Lemma 3 of Kveton et al. [2014])**.** *Let $\Delta_1 \geq \cdots \geq \Delta_K$ be a sequence of K positive numbers. Then,*

$$\sum_{k=1}^{K-1} (\Delta_k - \Delta_{k+1}) \frac{1}{\Delta_k^2} + \Delta_K \frac{1}{\Delta_K^2} = \Delta_1 \frac{1}{\Delta_1^2} + \sum_{k=2}^{K} \Delta_k \left(\frac{1}{\Delta_k^2} - \frac{1}{\Delta_{k-1}^2}\right) \leq \frac{2}{\Delta_K}. \tag{37}$$

Proof. First, we rewrite the expression as

$$\Delta_1 \frac{1}{\Delta_1^2} + \sum_{k=2}^{K} \Delta_k \left(\frac{1}{\Delta_k^2} - \frac{1}{\Delta_{k-1}^2} \right) = \sum_{k=1}^{K-1} \frac{\Delta_k - \Delta_{k+1}}{\Delta_k^2} + \frac{1}{\Delta_K}. \tag{38}$$

Next, by the assumption, we have $\Delta_k \geq \Delta_{k+1}$ for all $k < K$. Therefore,

$$\sum_{k=1}^{K-1} \frac{\Delta_k - \Delta_{k+1}}{\Delta_k^2} + \frac{1}{\Delta_K} \leq \sum_{k=1}^{K-1} \frac{\Delta_k - \Delta_{k+1}}{\Delta_k \Delta_{k+1}} + \frac{1}{\Delta_K} \tag{39}$$

$$= \sum_{k=1}^{K-1} \left(\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} \right) + \frac{1}{\Delta_K} = \frac{2}{\Delta_K} - \frac{1}{\Delta_1} < \frac{2}{\Delta_K}. \tag{40}$$

This completes the proof. □

Lemma 4. *Let*

$$x_1 \geq x_2 \geq \cdots \geq x_n > 0, \tag{41}$$

and let $a, b > 0$. Define

$$t_i := \begin{cases} 0 & \text{if } i = 0, \\ \min\left\{ \frac{a}{x_i^2},\, b \right\} & \text{if } i \neq 0. \end{cases} \tag{42}$$

Then,

$$\sum_{i=1}^{n-1} (x_i - x_{i+1}) t_i + x_n t_n = \sum_{i=1}^{n} x_i (t_i - t_{i-1}) \leq 2\sqrt{ab}. \tag{43}$$

Proof. Since $(x_i)_{i=1}^n$ is nonincreasing, the sequence $(a/x_i^2)_{i=1}^n$ is nondecreasing. Hence,

$$0 = t_0 \leq t_1 \leq \cdots \leq t_n \leq b. \tag{44}$$

We bound the left-hand side as

$$\sum_{i=1}^{n} x_i \left(t_i - t_{i-1} \right) \leq \sum_{i=1}^{n} \sqrt{a} \frac{t_i - t_{i-1}}{\sqrt{t_i}}, \tag{45}$$

where we used the fact that $t_i \leq a/x_i^2$, and thus $x_i \leq \sqrt{a}/\sqrt{t_i}$.

Now note that the function $f(t) = t^{-1/2}$ is decreasing on $(0, \infty)$. Therefore, for each interval $[t_{i-1}, t_i]$,

$$(t_i - t_{i-1}) \frac{1}{\sqrt{t_i}} \leq \int_{t_{i-1}}^{t_i} \frac{1}{\sqrt{t}}\, dt. \tag{46}$$

Summing over $i = 1, \ldots, n$, we obtain

$$\sum_{i=1}^{n} \sqrt{a} \frac{t_i - t_{i-1}}{\sqrt{t_i}} \leq \sqrt{a} \int_0^b \frac{1}{\sqrt{t}}\, dt = 2\sqrt{ab}. \tag{47}$$

Combining the above inequalities yields

$$\sum_{i=1}^{n} x_i \left(t_i - t_{i-1} \right) \leq 2\sqrt{ab}. \tag{48}$$

This completes the proof. □

D Proof of Theorems 1 and 2

Lemma 5 (Good-event per-round regret)**.** *Suppose the options are indexed so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_L$. For $j \in [(i-1) \wedge L^*]$, define*

$$n_{j,i,t} := \begin{cases} +\infty & \text{if } \rho_j = \rho_i, \\ \frac{6(1+\rho_j)^2 \log t}{(\rho_j - \rho_i)^2 c_i^2} & \text{if } \rho_j \neq \rho_i. \end{cases} \tag{49}$$

On the good event $\bar{\mathcal{E}}_t$,

$$\mathbb{E}_t[r(I^*)] - \mathbb{E}_t[r(I_t)] \leq \sum_{i=1}^{L} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}] \sum_{j=1}^{i-1 \wedge L^*} \mathbf{1}\{n_{j-1,i,t} < N_{i,t} \leq n_{j,i,t}\} c_i \left(1 - \frac{\rho_i}{\rho_j}\right) + \sum_{i \in [L] \setminus [L^*]} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}](c_i - \theta_i). \tag{50}$$

Proof. We first bound the loss from moving one option to its correct position in the efficiency order. Moving $I_t(i)$ from position i to position j is a sequence of adjacent swaps. For the r-th swap, the sample-path reward difference is

$$\delta_r = \mathbf{1}\{\mathcal{R}_t(i)\} \left(\prod_{q=1}^{r-1} (1 - X_t(I_t(i+q))) \right) \Big(Y_t(I_t(i)) X_t(I_t(i+r)) - Y_t(I_t(i+r)) X_t(I_t(i)) \Big). \tag{51}$$

Taking conditional expectation gives

$$\mathbb{E}_t[\delta_r] = \mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}] \prod_{q=1}^{r-1} (1 - \theta_{I_t(i+q)}) (c_{I_t(i)} \theta_{I_t(i+r)} - c_{I_t(i+r)} \theta_{I_t(i)}). \tag{52}$$

Thus the loss from moving an item is the sum of these adjacent-swap losses.

We now define the intermediate policies used in the telescoping argument. For an ordered list $I = (I(1), \ldots, I(|I|))$, let $\mathrm{OptSuffix}_i(I)$ be the ordered list obtained from the suffix $(I(i), I(i+1), \ldots, I(|I|))$ by first removing all options with efficiency strictly smaller than one, and then sorting the remaining options in descending order of efficiency, with ties broken according to the fixed global ordering. For each $i = 1, \ldots, |I_t| + 1$, define

$$I_t^{(i)} := (I_t(1), \ldots, I_t(i-1)) \| \, \mathrm{OptSuffix}_i(I_t), \tag{53}$$

where $\|$ denotes concatenation. By convention,

$$I_t^{(|I_t|+1)} := (I_t(1), \ldots, I_t(|I_t|)). \tag{54}$$

Thus, on the good event $\bar{\mathcal{E}}_t$, we have $I_t^{(1)} = I^*$ because all optimal options are included in I_t and the whole list is reordered optimally after removing inefficient options. Also, $I_t^{(|I_t|+1)} = I_t$.

For each i, the policies $I_t^{(i)}$ and $I_t^{(i+1)}$ have the same prefix up to position $i-1$. Hence their reward difference can occur only when the process reaches position i in the original list, namely on $\mathcal{R}_t(i)$. Equivalently, for option $I_t(i)$ this is the testing event $\mathcal{T}_{I_t(i),t}$.

The transition from $I_t^{(i+1)}$ to $I_t^{(i)}$ consists of moving the original option $I_t(i)$ into the optimally sorted suffix and, if $\rho_{I_t(i)} < 1$, removing it from the suffix. The loss from the moving operation is the sum of the adjacent-swap losses above. The removing operation is needed only when the moved option is inefficient. If the moved option is deleted after being moved behind positions $i+1, \ldots, j$, then the exact improvement from this deletion is

$$\mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}] \prod_{b=i+1}^{j} (1 - \theta_{I_t^{(i+1)}(b)}) (c_{I_t^{(i+1)}(i)} - \theta_{I_t^{(i+1)}(i)}) \mathbf{1}\{I_t^{(i+1)}(i) \in [L] \setminus [L^*]\}. \tag{55}$$

Indeed, after reaching the deleted option, keeping it would contribute expected single-option expected reward $\theta_{I_t^{(i+1)}(i)} - c_{I_t^{(i+1)}(i)}$, and the product is the probability that all options placed before it after the swaps fail. Since the product is at most one, the deletion improvement is upper bounded by the second term in the display below. Therefore, the preceding adjacent-swap calculation gives

$$\begin{aligned}
&\mathbb{E}_t[(r(I_t^{(i)}) - r(I_t^{(i+1)}))] \\
&\le \mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}] \sum_{a=i+1}^{j} \prod_{b=i+1}^{a-1} (1 - \theta_{I_t^{(i+1)}(b)})(\theta_{I_t^{(i+1)}(a)} c_{I_t^{(i+1)}(i)} - c_{I_t^{(i+1)}(a)} \theta_{I_t^{(i+1)}(i)}) \\
&\quad + \mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}](c_{I_t^{(i+1)}(i)} - \theta_{I_t^{(i+1)}(i)}) \mathbf{1}\{I_t^{(i+1)}(i) \in [L] \setminus [L^*]\}. \qquad (56)
\end{aligned}$$

For the swap term, use

$$\theta_{I_t^{(i+1)}(a)} c_{I_t^{(i+1)}(i)} - c_{I_t^{(i+1)}(a)} \theta_{I_t^{(i+1)}(i)} = \theta_{I_t^{(i+1)}(a)} c_{I_t^{(i+1)}(i)} \left(1 - \frac{\rho_{I_t^{(i+1)}(i)}}{\rho_{I_t^{(i+1)}(a)}}\right), \qquad (57)$$

and note that the largest efficiency among the options passed by $I_t^{(i+1)}(i)$ is $\rho_{I_t^{(i+1)}(i+1)}$. Since

$$\sum_{a=i+1}^{j} \prod_{b=i+1}^{a-1} (1 - \theta_{I_t^{(i+1)}(b)}) \theta_{I_t^{(i+1)}(a)} \le 1,$$

we obtain

$$\begin{aligned}
(56) &\le \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] c_{I_t(i)} \max_{k \in (\{I_t(j) | j > i\} \cap [L^*]) \cup \{I_t(i)\}} \left(1 - \frac{\rho_{I_t(i)}}{\rho_k}\right) \\
&\quad + \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}) \mathbf{1}\{I_t(i) \in [L] \setminus [L^*]\}. \qquad (58)
\end{aligned}$$

This also covers the case $j = i$.

By Lemma 1, as option $I_t(i)$ is sampled more often, the possible optimal options that can be misordered after it shrink according to the thresholds $n_{j,I_t(i),t}$. Hence

$$\begin{aligned}
(58) \le &\sum_{j=1}^{I_t(i)-1 \wedge L^*} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] \mathbf{1}\{n_{j-1,I_t(i),t} < N_{I_t(i),t} \le n_{j,I_t(i),t}\} c_{I_t(i)} \left(1 - \frac{\rho_{I_t(i)}}{\rho_j}\right) \\
&+ \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}) \mathbf{1}\{I_t(i) \in [L] \setminus [L^*]\}. \qquad (59)
\end{aligned}$$

On $\bar{\mathcal{E}}_t$, the set I_t contains all optimal options, so $I^* = I_t^{(1)}$. Also, if $\bar{\mathcal{E}}_t \cap \{n_{I_t(i)-1 \wedge L^*} < N_{I_t(i),t}\}$ holds, then no optimal option more efficient than $I_t(i)$ can appear after $I_t(i)$. A telescoping sum over the suffix-sorting sequence therefore yields

$$\begin{aligned}
&\mathbb{E}_t[r(I^*)] - \mathbb{E}_t[r(I_t)] \qquad (60) \\
&\le \sum_{i=1}^{|I_t|-1} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] \sum_{j=1}^{I_t(i)-1 \wedge L^*} \mathbf{1}\{n_{j-1,I_t(i),t} < N_{I_t(i),t} \le n_{j,I_t(i),t}\} c_{I_t(i)} \left(1 - \frac{\rho_{I_t(i)}}{\rho_j}\right) \\
&\quad + \sum_{I_t(i) \in [L] \setminus [L^*]} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}). \qquad (61)
\end{aligned}$$

Finally, reindex the sum over positions in I_t as a sum over option indices and extend it to all $i \in [L]$. If option i is not included in I_t, then $\mathcal{T}_{i,t}$ does not occur, so the extension adds zero terms. This gives (50). □

Lemma 6 (Bad-event per-round regret)**.** *The cumulative regret contribution on the bad events satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t[(r(I^*) - r(I_t))] \mathbf{1}\{\mathcal{E}_t\}\right] \le 2\left(1 + \frac{4\pi^2}{3}\right) L^2. \qquad (62)$$

Proof. At each round, the regret is at most $2L$. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t[(r(I^*)-r(I_t))]\mathbf{1}\{\mathcal{E}_t\}\right] \le \mathbb{E}\left[\sum_{t=1}^{T}|\mathbb{E}_t[(r(I^*)-r(I_t))]|\,\mathbf{1}\{\mathcal{E}_t\}\right]$$
$$\le \mathbb{E}\left[2L\sum_{t=1}^{T}\mathbf{1}\{\mathcal{E}_t\}\right]. \tag{63}$$

By Lemma 2 of Zhou et al. [2018], $\mathbb{E}[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{E}_t\}] \le (1+4\pi^2/3)L$, which proves (62). □

Proof. Reindex the options in descending order of efficiency, so that $\rho_1 \ge \rho_2 \ge \cdots \ge \rho_L$. We first isolate the regret on the good and bad events. By the tower property,

$$\mathbb{E}[\sum_{t=1}^{T}(r(I^*)-r(I_t))] \tag{64}$$
$$= \mathbb{E}[\sum_{t=1}^{T}\mathbb{E}_t[(r(I^*)-r(I_t))]] \tag{65}$$
$$= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t[(r(I^*)-r(I_t))]\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_t[(r(I^*)-r(I_t))]\mathbf{1}\{\mathcal{E}_t\}\right]. \tag{66}$$

Applying Lemmas 5 and 6 to the two terms gives

$$(66) \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{L}\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\sum_{j=1}^{i-1\wedge L^*}\mathbf{1}\{n_{j-1,i,t} < N_{i,t} \le n_{j,i,t}\}c_i\left(1-\frac{\rho_i}{\rho_j}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]}\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}](c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2. \tag{67}$$

It remains to rewrite the first term in a form that separates adjacent efficiency gaps. Since

$$\mathbf{1}\{n_{j-1,i,t} < N_{i,t} \le n_{j,i,t}\} = \mathbf{1}\{N_{i,t} \le n_{j,i,t}\} - \mathbf{1}\{N_{i,t} \le n_{j-1,i,t}\},$$

summation by parts yields the common bound

$$\mathbb{E}[\sum_{t=1}^{T}(r(I^*)-r(I_t))] \tag{68}$$
$$\le \sum_{i=1}^{L}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\sum_{j=1}^{i-1\wedge L^*}\left(\mathbf{1}\{N_{i,t}\le n_{j,i,t}\}-\mathbf{1}\{N_{i,t}\le n_{j-1,i,t}\}\right)c_i\left(1-\frac{\rho_i}{\rho_j}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{69}$$
$$= \sum_{i=1}^{L}\mathbb{E}\left[\sum_{j=1}^{i-1\wedge L^*}\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\left(\mathbf{1}\{N_{i,t}\le n_{j,i,t}\}-\mathbf{1}\{N_{i,t}\le n_{j-1,i,t}\}\right)c_i\left(1-\frac{\rho_i}{\rho_j}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{70}$$
$$= \sum_{i=1}^{L}\mathbb{E}\left[\sum_{j=1}^{i-1\wedge L^*-1}\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{j,i,t}\}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$

$$+\sum_{i=1}^{L}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{i-1\wedge L^*,i,t}\}c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]+2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{71}$$

We now derive the two regret bounds from (71).

Case 1: Problem-dependent bound We split (71) into three contributions: misordering optimal options, misordering suboptimal options, and selecting suboptimal options. The suboptimal terms are simpler, so we bound them first.

Misordering suboptimal options. Fix $i\in[L]\setminus[L^*]$. Since $n_{j,i,t}\le n_{j,i,T}$,

$$\sum_{j=1}^{i-1\wedge L^*-1}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{j,i,t}\}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{i-1\wedge L^*,i,t}\}c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] \tag{72}$$
$$\le\sum_{j=1}^{i-1\wedge L^*-1}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{j,i,T}\}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{i-1\wedge L^*,i,T}\}c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] \tag{73}$$

Moreover,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{j,i,T}\}\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]\le n_{j,i,T},$$

because $N_{i,t}$ increases by one whenever option i is tested, and option i is tested at most once at each round. Therefore,

$$(72)\le\sum_{j=1}^{i-1\wedge L^*-1}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\frac{6(1+\rho_j)^2\log T}{(\rho_j-\rho_i)^2c_i^2}$$
$$+c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\frac{6(1+\rho_{i-1\wedge L^*})^2\log T}{(\rho_{i-1\wedge L^*}-\rho_i)^2c_i} \tag{74}$$
$$\le\sum_{j=1}^{i-1\wedge L^*-1}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\frac{24\log T}{\left(1-\frac{\rho_i}{\rho_j}\right)^2c_i^2}$$
$$+c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\frac{24\log T}{\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)^2c_i}\qquad(\because\rho_j\ge1\text{ for }j\le L^*) \tag{75}$$
$$\le\frac{48\log T}{(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}})c_i}\qquad(\because\text{ Lemma 3}) \tag{76}$$
$$\le\frac{48\log T}{(1-\rho_i)c_i}=\frac{48\log T}{c_i-\theta_i}. \tag{77}$$

Selecting suboptimal options. From Lemma 2, the number of observations of the suboptimal option i under $\bar{\mathcal{E}}_t$ is bounded from above by $\frac{24\log T}{(c_i-\theta_i)^2}$. Hence,

$$\sum_{i\in[L]\setminus[L^*]}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]\le\sum_{i\in[L]\setminus[L^*]}\frac{24\log T}{c_i-\theta_i}. \tag{78}$$

Misordering optimal options. For each optimal option i, define

$$p_i := \frac{\prod_{j=1}^{L}(1-\theta_j)}{(1-\theta_i)}, \tag{79}$$

$$\zeta_{j,i} := \max\left\{ n : \frac{6(1+\rho_j)^2(\log n + \log 2)}{(\rho_j-\rho_i)^2 c_i^2} \geq \frac{p_i}{2} n \right\}, \tag{80}$$

and

$$\eta_0 := 16L\left(\frac{\pi^2}{6} + 1 + \log 2 + \frac{1}{3}(2 + \log^2 3 + 2\log 3)\right). \tag{81}$$

We also define the closest strictly more efficient optimal option above i. Let

$$S_i := \{\, j \in [L^*] \mid \rho_j - \rho_i > 0,\ j < i \,\}, \tag{82}$$

$$\Delta_{i,\min} := \begin{cases} \min_{j\in S_i}(\rho_j - \rho_i), & \text{if } S_i \neq \varnothing, \\ 0, & \text{if } S_i = \varnothing, \end{cases} \tag{83}$$

and let $\mathrm{idx}_{i,\min}$ be the maximum index satisfying $\rho_{\mathrm{idx}_{i,\min}} - \rho_i = \Delta_{i,\min}$ when $\Delta_{i,\min} > 0$.

Here we use the additional condition $\theta_i < 1$ for all options, which ensures that the reachability probabilities appearing in Lemma 3 of Zhou et al. [2018] are positive. The proof of that lemma then shows that, for any optimal option i,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\bar{\mathcal{E}}_t\}\mathbf{1}\{N_{i,t} \leq n_{i-1,i,t}\}\right] \leq \zeta_{i-1,i} + \frac{2}{p_i^2} + \eta_0. \tag{84}$$

The same argument extends to any $j < i-1$:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\bar{\mathcal{E}}_t\}\mathbf{1}\{N_{i,t} \leq n_{j,i,t}\}\right] \leq \zeta_{j,i} + \frac{2}{p_i^2} + \eta_0. \tag{85}$$

Therefore the optimal-option misordering term in (71) satisfies

$$\begin{aligned}
&\sum_{i=1}^{L^*}\sum_{j=1}^{i-1\wedge L^*-1} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t} \leq n_{j,i,t}\}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right) - c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] \\
&\quad + \sum_{i=1}^{L^*} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t} \leq n_{i-1\wedge L^*,i,t}\} c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] \qquad (86) \\
&\leq \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \sum_{j=1}^{\mathrm{idx}_{i,\min}-1} \left(c_i\left(1-\frac{\rho_i}{\rho_j}\right) - c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\left(\zeta_{j,i} + \frac{2}{p_i^2} + \eta_0\right) \\
&\quad + \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \left(c_i\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)\right)\left(\zeta_{\mathrm{idx}_{i,\min},i} + \frac{2}{p_i^2} + \eta_0\right) \qquad (87) \\
&\leq \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \sum_{j=1}^{\mathrm{idx}_{i,\min}-1} \left(c_i\left(1-\frac{\rho_i}{\rho_j}\right) - c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\zeta_{j,i} \\
&\quad + \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \left(c_i\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)\right)\zeta_{\mathrm{idx}_{i,\min},i} + \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} c_i\left(\frac{2}{p_i^2} + \eta_0\right) \qquad (88) \\
&\leq \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \sum_{j=1}^{\mathrm{idx}_{i,\min}-1} \left(c_i\left(1-\frac{\rho_i}{\rho_j}\right) - c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\frac{48(\log\zeta_{j,i} + \log 2)}{\left(1-\frac{\rho_i}{\rho_j}\right)^2 c_i^2 p_i}
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \left(c_i\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)\right)\frac{48(\log\zeta_{\mathrm{idx}_{i,\min},i}+\log 2)}{\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)^2 c_i^2 p_i}\\
&+ \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} c_i\left(\frac{2}{p_i^2}+\eta_0\right) \qquad (\because \text{(80) and } \rho_j\geq 1 \text{ for } j\leq L^*) \qquad (89)\\
\leq &\sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \sum_{j=1}^{\mathrm{idx}_{i,\min}-1} \left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\frac{48(\log\zeta_{\mathrm{idx}_{i,\min},i}+\log 2)}{\left(1-\frac{\rho_i}{\rho_j}\right)^2 c_i^2 p_i}\\
&+ \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \left(c_i\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)\right)\frac{48(\log\zeta_{\mathrm{idx}_{i,\min},i}+\log 2)}{\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)^2 c_i^2 p_i}\\
&+ \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} c_i\left(\frac{2}{p_i^2}+\eta_0\right) \qquad (90)\\
\leq &\sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \frac{96(\log\zeta_{\mathrm{idx}_{i,\min},i}+\log 2)}{\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)c_i p_i} + \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} c_i\left(\frac{2}{p_i^2}+\eta_0\right) \qquad (\because \text{Lemma 3}). \qquad (91)
\end{aligned}
$$

Combining the three bounds (91), (77), and (78) in (71), we obtain

$$
\begin{aligned}
(71) \leq &\sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} \frac{96(\log\zeta_{\mathrm{idx}_{i,\min},i}+\log 2)}{\left(1-\frac{\rho_i}{\rho_{\mathrm{idx}_{i,\min}}}\right)c_i p_i} + \sum_{\substack{i\in[L^*]\\ \Delta_{i,\min}>0}} c_i\left(\frac{2}{p_i^2}+\eta_0\right)\\
&+ \sum_{i\in[L]\setminus[L^*]} \frac{72\log T}{c_i-\theta_i} + 2\left(1+\frac{4\pi^2}{3}\right)L^2. \qquad (92)
\end{aligned}
$$

Case 2: Problem-independent bound For the problem-independent bound, we truncate each threshold at the horizon. Since $N_{i,t}\leq T$,

$$
\mathbf{1}\{N_{i,t}\leq n_{j,i,t}\}\leq \mathbf{1}\{N_{i,t}\leq n_{j,i,T}\}\leq \mathbf{1}\{N_{i,t}\leq \min\{n_{j,i,T},T\}\}. \qquad (93)
$$

Hence,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\leq n_{j,i,t}\}\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\leq \min\{n_{j,i,T},T\}\}\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] \qquad (94)\\
&\leq \min\{n_{j,i,T},T\} = \min\{\frac{6(1+\rho_j)^2\log T}{(\rho_j-\rho_i)^2 c_i^2},T\}. \qquad (95)
\end{aligned}
$$

Substituting this bound into (71) gives

$$
\begin{aligned}
(71) \leq &\sum_{i=1}^{L}\sum_{j=1}^{i-1\wedge L^*-1}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\min\left\{\frac{6(1+\rho_j)^2\log T}{(\rho_j-\rho_i)^2c_i^2},T\right\}\\
&+\sum_{i=1}^{L}c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\min\left\{\frac{6(1+\rho_{i-1\wedge L^*})^2\log T}{(\rho_{i-1\wedge L^*}-\rho_i)^2c_i^2},T\right\}\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]+2\left(1+\frac{4\pi^2}{3}\right)L^2 \qquad (96)\\
\leq &\sum_{i=1}^{L}\sum_{j=1}^{i-1\wedge L^*-1}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right)-c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\min\left\{\frac{24\log T}{(1-\frac{\rho_i}{\rho_j})^2c_i^2},T\right\}
\end{aligned}
$$

$$+\sum_{i=1}^{L} c_i \left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right) \min\left\{\frac{24\log T}{(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}})^2 c_i^2}, T\right\}$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]} \mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \qquad (\because \rho_j \geq 1 \text{ for } j \leq L^*) \tag{97}$$

$$\leq 2L\sqrt{24T\log T}$$
$$+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]} \mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \qquad (\because \text{Lemma 4}) \tag{98}$$

$$\leq 2L\sqrt{24T\log T} + \sum_{i\in[L]\setminus[L^*]} \mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{99}$$

$$\leq 2L\sqrt{24T\log T}$$
$$+\sum_{i\in[L]\setminus[L^*]} \min\left\{\frac{24\log T}{(c_i-\theta_i)^2}, T\right\}(c_i-\theta_i) + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \qquad (\because \text{Lemma 2}) \tag{100}$$

$$\leq 2L\sqrt{24T\log T} + (L-L^*)\sqrt{24T\log T} + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{101}$$

□

E Proof of Theorems 4

Proof. In this proof, we reindex the option set in descending order of efficiency, so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_L$. Also, L' is the number of options whose efficiency is ∞. We show that the inequality (58) still applies when the option set $[L]$ contains zero-cost options.

Recall that, for CC-UCBv2, we define

$$\rho_i := \begin{cases} \dfrac{\theta_i}{c_i}, & \text{if } c_i > 0, \\ +\infty, & \text{if } c_i = 0 \wedge \theta_i > 0, \end{cases} \tag{102}$$

so zero-cost options with positive success probability have efficiency $+\infty$. Throughout this proof, ratios involving $+\infty$ are interpreted in the extended-real sense: if $\rho_i < +\infty$, then $\rho_i/(+\infty) = 0$. Options with $c_i = \theta_i = 0$ never produce a success and incur no cost, so including or ordering them does not change the expected reward of any list. We therefore ignore these options in the proof; if they remain in the algorithmic list, they do not contribute to regret. It remains to consider the case where the options passed by $I_t^{(i+1)}(i)$ include such a zero-cost option. In this case, the efficiency-gap expression used in (58) is equal to one for that passed option, and we use the following coarse bound on the swap loss:

$$(56) \leq \mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}] \sum_{a=i+1}^{j}\prod_{b=i+1}^{a-1}(1-\theta_{I_t^{(i+1)}(b)})(\theta_{I_t^{(i+1)}(a)}c_{I_t^{(i+1)}(i)} - c_{I_t^{(i+1)}(a)}\theta_{I_t^{(i+1)}(i)})$$
$$+\mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}](c_{I_t^{(i+1)}(i)} - \theta_{I_t^{(i+1)}(i)})\mathbf{1}\{I_t^{(i+1)}(i) \in [L]\setminus[L^*]\} \tag{103}$$

$$\leq \mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}] \sum_{a=i+1}^{j}\prod_{b=i+1}^{a-1}(1-\theta_{I_t^{(i+1)}(b)})\theta_{I_t^{(i+1)}(a)}c_{I_t^{(i+1)}(i)}$$
$$+\mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}](c_{I_t^{(i+1)}(i)} - \theta_{I_t^{(i+1)}(i)})\mathbf{1}\{I_t^{(i+1)}(i) \in [L]\setminus[L^*]\} \tag{104}$$

$$\leq \mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}]c_{I_t^{(i+1)}(i)}$$
$$+\mathbb{E}_t[\mathbf{1}\{\mathcal{R}_t(i)\}](c_{I_t^{(i+1)}(i)} - \theta_{I_t^{(i+1)}(i)})\mathbf{1}\{I_t^{(i+1)}(i) \in [L]\setminus[L^*]\}. \tag{105}$$

This is exactly the value of the first term in (58) when the maximum is attained by a zero-cost option with efficiency $+\infty$. Therefore, the same per-round decomposition used in Appendix D remains valid for CC-UCBv2.

Also, we redefine $n_{j,i,t}$ as in Lemma 1:

$$n_{j,i,t} := \begin{cases} +\infty & \text{if } \rho_j = \rho_i, \\ \frac{6\log t}{c_i^2} & \text{if } c_j = 0 \text{ and } \rho_j \neq \rho_i, \\ \frac{6(1+\rho_j)^2 \log t}{(\rho_j - \rho_i)^2 c_i^2} & \text{otherwise.} \end{cases} \tag{106}$$

Lemma 1 then implies that, on $\bar{\mathcal{E}}_t$, if $n_{j-1,i,t} < N_{i,t} \leq n_{j,i,t}$, no option more efficient than j can appear after i under the CC-UCBv2 ordering rule. This includes the case $c_j = 0$, where the threshold $6\log t/c_i^2$ prevents a zero-cost option from being misordered after a positive-cost option once $L_{i,t} > 0$.

We now repeat the problem-independent summation argument from Appendix D, but only over finite-efficiency options. This restriction is valid because the first L' options have efficiency $+\infty$: they are optimal, and no more efficient option can be misordered after them. This gives

$$\mathbb{E}[\sum_{t=1}^{T}(r(I^*) - r(I_t))] \tag{107}$$

$$\leq \sum_{i=L'+1}^{L} \mathbb{E}\left[\sum_{j=1}^{i-1\wedge L^*-1}\sum_{t=1}^{T} \mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t} \leq n_{j,i,t}\}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right) - c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+ \sum_{i=L'+1}^{L} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t} \leq n_{i-1\wedge L^*,i,t}\}c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+ \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]} \mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{108}$$

$$\leq \sum_{i=L'+1}^{L}\sum_{j=1}^{i-1\wedge L^*-1} c_i\left(1-\frac{\rho_i}{\rho_j}\right)\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t} \leq \min\{n_{j,i,T},T\}\}\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$- \sum_{i=L'+1}^{L}\sum_{j=1}^{i-1\wedge L^*-1} c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t} \leq \min\{n_{j,i,T},T\}\}\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+ \sum_{i=L'+1}^{L} c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t} \leq \min\{n_{i-1\wedge L^*,i,T},T\}\}\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]$$
$$+ \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]} \mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{109}$$

$$\leq \sum_{i=L'+1}^{L}\sum_{j=1}^{i-1\wedge L^*-1}\left(c_i\left(1-\frac{\rho_i}{\rho_j}\right) - c_i\left(1-\frac{\rho_i}{\rho_{j+1}}\right)\right)\min\left\{\frac{24\log T}{(1-\frac{\rho_i}{\rho_j})^2 c_i^2}, T\right\}$$
$$+ \sum_{i=L'+1}^{L} c_i\left(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}}\right)\min\left\{\frac{24\log T}{(1-\frac{\rho_i}{\rho_{i-1\wedge L^*}})^2 c_i^2}, T\right\}$$
$$+ \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]} \mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right] + 2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{110}$$

$$\leq 2(L-L')\sqrt{24T\log T}$$

$$+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in[L]\setminus[L^*]}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]+2\left(1+\frac{4\pi^2}{3}\right)L^2 \qquad (\because \text{Lemma 4}) \tag{111}$$

$$\leq 2(L-L')\sqrt{24T\log T}+\sum_{i\in[L]\setminus[L^*]}\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}(c_i-\theta_i)\mathbf{1}\{\bar{\mathcal{E}}_t\}\right]+2\left(1+\frac{4\pi^2}{3}\right)L^2 \tag{112}$$

$$\leq 2(L-L')\sqrt{24T\log T}$$
$$+\sum_{i\in[L]\setminus[L^*]}\min\left\{\frac{24\log T}{(c_i-\theta_i)^2},T\right\}(c_i-\theta_i)+2\left(1+\frac{4\pi^2}{3}\right)L^2 \qquad (\because \text{Lemma 2}) \tag{113}$$

$$\leq 2(L-L')\sqrt{24T\log T}+(L-L^*)\sqrt{24T\log T}+2\left(1+\frac{4\pi^2}{3}\right)L^2. \tag{114}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the theoretical bounds, the CC-UCBv2 extension, and the scope of the numerical experiments. These claims are supported by the stated theorems, proofs in the appendix, and experiments in the numerical section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a separate Limitations section. It discusses the remaining gap between upper and lower bounds, the limitations of UCB-style confidence intervals, and the illustrative scope of the numerical experiments, which focus on phenomena tied to the theoretical and algorithmic questions of the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are stated in the problem setting and algorithm sections, and each theorem is accompanied by either a proof sketch in the main text or a full proof in the appendix. The auxiliary lemmas used in the proofs are stated and referenced in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The numerical experiments section specifies the bandit instances, horizon, number of independent simulations, and compared algorithms. The anonymized supplementary code includes a reproduction notebook and README with the environment and commands needed to regenerate the reported figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide anonymized supplementary code with independent implementations of the algorithms and a notebook for reproducing the reported simulation figures. The supplementary README specifies the required environment, commands, and output files.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: The numerical experiments section specifies the Bernoulli instances, horizons, number of independent simulations, compared algorithms, and values of the lower-bound parameter used by CC-UCB. The supplementary code provides the exact reproduction notebook and implementation details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The numerical experiments section states that curves show the mean cumulative regret over independent simulations and error bars show one empirical standard deviation across runs. The supplementary reproduction code computes these quantities directly from the independent simulation runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The numerical experiments section states that all experiments are lightweight CPU-based simulations and require no GPU or specialized hardware. It also specifies the horizon and number of independent simulations used for each reported figure.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is a theoretical and simulation-based study and does not involve human subjects, private data, scraped data, or high-risk model release. The supplementary code is anonymized for review.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Limitations section discusses possible broader impacts of cost-aware sequential testing, including potential cost reductions in applied systems and risks from using the model in high-stakes domains without validating assumptions. The paper is primarily theoretical and does not deploy a system or use sensitive data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: We implement CC-UCB ourselves based on the algorithm described in prior work, which is cited in the paper. We do not reuse or redistribute existing code, datasets, or other assets from that work.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released supplementary code is documented with a README, dependency file, reproduction notebook, helper script, and license. It is anonymized for review and contains no datasets, models, or human-subject data.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as part of the core methodology, theoretical results, algorithms, or experiments. Any writing or editing assistance does not affect the scientific content or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.