Regret Decomposition and 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 inefficiency 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 $\tilde{O}(L\sqrt{T})$. These bounds assume a known positive lower bound on the costs, and the problem-dependent bound further assumes that there is no sure-success option, that is, no option that succeeds with certainty. We then remove these assumptions one at a time. CC-UCBv2 removes the need to specify a positive lower bound on costs and handles zero-cost options while preserving both bounds. BC-CC-UCBv2 further removes the assumption on sure-success options by certifying the stopping boundary induced by such options, and admits a problem-dependent bound of $O(\sum_{i:c_i>c^*\theta_i} \log T/(c_i - c^*\theta_i))$, where c^* is the smallest cost of a sure-success option ($c^* = 1$ if none exists). We also 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. Numerical experiments illustrate the effect of a misspecified cost lower bound on efficiency estimates and compare the behavior of the modified algorithms with the CC-UCB baseline.

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., 2020b]. 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

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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

Paper	Result	Assumptions	Problem Indep.	Problem Dependent
Zhou et al. [2018]	CC-UCB	ϵ	none	$O(\sum_{i:\rho_i<1} c_i \frac{\log T}{(c_i-\theta_i)^2})^{\dagger}$
Zhou et al. [2018]	Lower bound	$\theta_i < 1$	none	$\Omega\Big(\sum_{i:\rho_i<1} \frac{(c_i-\theta_i)\log T}{d(\theta_i;c_i)}\Big)^{\S}$
Ours	CC-UCB	ϵ; $\theta_i < 1$ (PD)	$\tilde{O}(L\sqrt{T})$ (Th. 2)	$O(\sum_{i:\rho_i<1} \frac{\log T}{c_i-\theta_i})^{\dagger}$ (Th. 1)
Ours	CC-UCBv2	$\theta_i < 1$ (PD)	$\tilde{O}(L\sqrt{T})$ (Th. 3)	$O(\sum_{i:\rho_i<1} \frac{\log T}{c_i-\theta_i})^{\dagger}$ (Sec. 5.1)
Ours	BC-CC-UCBv2	none	none	$O(\sum_{i:c_i>c^*\theta_i} \frac{\log T}{c_i-c^*\theta_i})^{\dagger\ddagger}$ (Th. 4)
Ours	Lower bound	—	$\Omega(\sqrt{LT})$ (Th. 5)	none

ϵ: a known positive lower bound on the expected costs. PD: problem-dependent bound.
† This upper bound ignores terms that do not depend on T.
‡ c^* is the smallest cost of a sure-success option ($\theta_i = 1$), and $c^* = 1$ if there is no such option. When $\theta_i < 1$ for all i, the bound reduces to that of CC-UCB.
§ $d(\theta_i; c_i)$ is the KL divergence of Bernoulli distributions with means θ_i and c_i. The condition $\theta_i < 1$ is not stated by Zhou et al. [2018], but without it, an option with $\theta_j = 1$ can prevent later options from ever being observed.

random time, signaling, or energy cost $Y_{i,t}$ with mean c_i. Some candidates may satisfy the condition with certainty. For example, a fallback cell that always accepts the connection but at a higher signaling or energy cost corresponds to a sure-success option, that is, an option with $\theta_i = 1$; once it is tested, the examination stops. 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 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 $I_t = (I_t(1), \dots, I_t(|I_t|))$ of options and tests them sequentially. The process terminates once a success is observed, so the set of actually tested options is random. The reward r_t is the success indicator minus the total cost of the tested options (Section 3). 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., 2015b,a]. 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 a learning algorithm by its regret R_T, the gap between its expected cumulative reward over T rounds and that of an optimal ordered subset I^* chosen with knowledge of the true parameters (Section 3). 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 inefficiency gap $c_i - \theta_i$.

In this work, we show that the regret of CC-UCB admits a problem-dependent bound of $O\Big(\sum_{i\in[L]:\rho_i<1} \frac{\log T}{c_i-\theta_i}\Big)$ (Theorem 1). This improves the bound of Zhou et al. [2018], which is of order $O(\sum_{i\in[L]:\rho_i<1} \frac{c_i \log T}{(c_i-\theta_i)^2})$. 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 $\Omega\Big(\sum_{i\in[L]:\rho_i<1} \frac{(c_i-\theta_i)\log T}{d(\theta_i;c_i)}\Big)$, 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 $\Omega(\sum_{i\in[L]:\rho_i<1} \frac{\log T}{c_i-\theta_i})$. 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 5). 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 increases the expected reward by $\prod_{j=1}^{i-1}(1-\theta_{I_t(j)})(\theta_{I_t(i+1)}c_{I_t(i)} - c_{I_t(i+1)}\theta_{I_t(i)}) \geq 0$ (see (5) in Section 3). Directly comparing I_t with I^* 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. Each difference is bounded by two terms (Section 4). The first is the loss from testing $I_t(i)$ before a more efficient remaining optimal option. As $I_t(i)$ is sampled more often, such a misordering can occur only with options whose efficiency is closer to that of $I_t(i)$, so this loss diminishes. The second is the per-test loss $c_{I_t(i)} - \theta_{I_t(i)}$ when $I_t(i)$ is inefficient. Thus the decomposition separates the misordering loss from the regret of testing inefficient options.

The guarantees for CC-UCB rely on two assumptions: a known positive lower bound ϵ on the expected costs, and, for the problem-dependent bound, $\theta_i < 1$ for all i. We remove them one at a time. First, we propose CC-UCBv2, which does not require ϵ. Prior knowledge of such a lower bound is often unavailable in practice. CC-UCBv2 also handles zero-cost options and preserves both guarantees of CC-UCB (Section 5.1). Second, we propose BC-CC-UCBv2 (Boundary-Certifying CC-UCBv2), which also removes the assumption $\theta_i < 1$. When sure-success options exist, the optimal list stops at the cheapest such option, whose cost we denote by c^*, and the loss of testing option i becomes the boundary gap $c_i - c^*\theta_i$. The learner has to locate this stopping boundary without knowing c^*. Moreover, observations cannot confirm that an option is a sure-success option. A single failure shows that $\theta_i < 1$, but no number of successes shows that $\theta_i = 1$. BC-CC-UCBv2 therefore keeps options that have never failed at the end of the list and estimates c^* from them. Its regret admits a problem-dependent bound of $O(\sum_{i:c_i>c^*\theta_i} \log T/(c_i - c^*\theta_i))$ (Theorem 4), which reduces to the order of Theorem 1 when $\theta_i < 1$ for all i. Numerical experiments examine the distortion caused by a misspecified cost lower bound and compare the modified algorithms with CC-UCB in an instance containing zero-cost and sure-success options.

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. [2013] 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.

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., 2015b,a] 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 $w_1, \ldots, w_L$, 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., 2015a], 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. [2015a] 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.

Order-dependent expected rewards also arise in stochastic online learning to rank. TopRank [Lattimore et al., 2018] learns rankings under a general click model encompassing position-based and cascade models. GRAB [Gauthier et al., 2021] exploits unimodality on a learned graph of rankings in the position-based setting, while UniRank [Gauthier et al., 2022] uses a related graph structure to handle a broader class of click models. Safety during exploration has also been studied: BubbleRank [Li et al., 2020a] gradually improves an initial ranking while controlling the number of incorrectly ordered item pairs, and Shiino et al. [2023] extend safe re-ranking to exploration of items outside the displayed list. These studies focus on maximizing clicks, whereas CCB additionally accounts for the costs incurred by sequentially testing individual options until the first success.

Cost-aware cascading bandits (CCB) [Zhou et al., 2018] extend the standard CB 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. As in the introduction, the efficiency of an option with $c_i > 0$ is $\rho_i := \theta_i / c_i$. 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 5.1, where we use an extended-real definition of efficiency, and options with $\theta_i = 1$ are handled in Section 5.2. 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

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_L, \tag{1}$$

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 $I_t = (I_t(1), I_t(2), \ldots, I_t(|I_t|))$ 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,

$$r_t = 1 - \prod_{i=1}^{|\tilde{I}_t|}\big(1 - X_{\tilde{I}_t(i),t}\big) - \sum_{i=1}^{|\tilde{I}_t|} Y_{\tilde{I}_t(i),t}. \tag{2}$$

Equivalently, the reward can be written as

$$r_t = \sum_{k=1}^{|I_t|}\left(\prod_{j=1}^{k-1}\big(1 - X_{I_t(j),t}\big)\right)\Big(X_{I_t(k),t} - Y_{I_t(k),t}\Big). \tag{3}$$

We write $r_t(I)$ for the reward at round t when the ordered subset I is selected. By independence, the expected reward of a fixed ordered subset $I = (I(1), \ldots, I(n))$ is

$$\mu(I) := \sum_{k=1}^{n}\Big(\prod_{j=1}^{k-1}\big(1 - \theta_{I(j)}\big)\Big)\big(\theta_{I(k)} - c_{I(k)}\big). \tag{4}$$

Let $\mathcal{F}_t$ be the σ-algebra generated by all observations made before round t, and write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Since the algorithms in this paper choose I_t from these observations, I_t is $\mathcal{F}_t$-measurable and $\mathbb{E}_t[r_t(I_t)] = \mu(I_t)$.

We also call an ordered subset a list and write $I \| I'$ for the concatenation of lists I and I'. For lists $I = I_1 \| (i,j) \| I_2$ and $I' = I_1 \| (j,i) \| I_2$,

$$\mu(I') - \mu(I) = \Big(\prod_{k \in I_1}(1 - \theta_k)\Big)\big(\theta_j c_i - \theta_i c_j\big), \tag{5}$$

because, after the pair is reached, the expected net reward of (i,j) followed by I_2 is $\theta_i - c_i + (1 - \theta_i)(\theta_j - c_j) + (1-\theta_i)(1-\theta_j)\mu(I_2)$ and the last term is symmetric in i and j.

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

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\big(r_t(I^*) - r_t(I_t)\big)\right]. \tag{6}$$

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. Hence $[L^*]$ is the set of options with $\rho_i \geq 1$, and $[L] \setminus [L^*] = \{i \in [L] : \rho_i < 1\}$. We write $\mathrm{reg}_t := \mu(I^*) - \mu(I_t)$ for the conditional expected regret of round t; since I_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[r_t(I^*) - r_t(I_t)] = \mathbb{E}[\mathrm{reg}_t]$.

4 Analysis of CC-UCB

This section analyzes the original CC-UCB algorithm under the positive-cost assumption stated below. We define the following two events. The first concerns the sequential testing process, and the second concerns estimation errors of sample means.

Definition 1 (Testing event)**.** *For each option i, define*

$$\mathcal{T}_{i,t} := \{i \in I_t\} \cap \big\{X_{j,t} = 0 \ \ \forall j \in I_t \text{ s.t. } I_t^{-1}(j) < I_t^{-1}(i)\big\}. \tag{7}$$

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

Definition 2 (Error event)**.** *At round t, define the error event $\mathcal{E}_t$ as*

$$\mathcal{E}_t := \Big\{\exists i \in [L],\ \big|\hat{\theta}_{i,t} - \theta_i\big| > u_{i,t} \text{ or } \big|\hat{c}_{i,t} - c_i\big| > u_{i,t}\Big\}, \qquad u_{i,t} := \sqrt{\frac{1.5\log t}{N_{i,t}}}, \tag{8}$$

where $N_{i,t}$ is the number of observations of option i before round t, and $\hat{\theta}_{i,t}$ and $\hat{c}_{i,t}$ are the corresponding sample means. We call $\mathcal{E}_t^c$ the good event; on it, all estimates are sufficiently close to their true values.

4.1 Cost-Aware Cascading UCB (CC-UCB)

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 $\hat{c}_{i,t}$. Following Kveton et al. [2015a] and Kveton et al. [2015b], we assume for simplicity that one initial observation $(X_{i,0}, Y_{i,0})$ of each option i is available before round 1, and the algorithm initializes these quantities from it, so that $N_{i,1} = 1$. Such observations can be generated in L rounds by testing each option alone, that is, with $I_t = (i)$. Each of these rounds contributes at most 2 to the regret, since $\mu(I^*) \leq 1$ and $\theta_i - c_i \geq -1$, so generating the initial observations adds at most $2L$ to the regret. The same applies to CC-UCBv2 and BC-CC-UCBv2. At each round t, the algorithm uses the confidence radius $u_{i,t}$ defined above. 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

$$U_{i,t} = \hat{\theta}_{i,t} + u_{i,t}, \qquad L_{i,t} = \max\{\hat{c}_{i,t} - u_{i,t}, \epsilon\}. \tag{9}$$

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

$$\frac{U_{i,t}}{L_{i,t}} \geq 1. \tag{10}$$

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.

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*

$$R_T \leq \sum_{i \in [L]: \rho_i < 1} \frac{72 \log T}{c_i - \theta_i} + O(1). \tag{11}$$

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

The condition $\theta_i < 1$ ensures that every option is reached with positive probability even when other options precede it. Lemma 3 of Zhou et al. [2018] bounds how often optimal options are misordered (Lemma 8 in Appendix C). Its argument uses this property without stating the condition.

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

$$R_T \leq \tilde{O}(L\sqrt{T}). \tag{12}$$

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

Good-event reduction. The proof first separates the regret on the error event $\mathcal{E}_t$ from the regret on the good event $\mathcal{E}_t^c$. 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 $\mathcal{E}_t^c$.

Per-round decomposition. On the good event, the list I_t selected by CC-UCB contains every option in I^*. To compare I_t with I^*, we repeatedly sort suffixes and delete inefficient options. Applying (5)

to each step (Appendix C) yields

$$\mathrm{reg}_t \le \sum_{i=1}^{|I_t|-1} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] c_{I_t(i)} \max_{\substack{k \in \{I_t(i'):i'>i\} \cap [L^*] \\ \rho_k > \rho_{I_t(i)}}} \left(1 - \frac{\rho_{I_t(i)}}{\rho_k}\right) + \sum_{\substack{1 \le i \le |I_t| \\ I_t(i) \in [L] \setminus [L^*]}} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}). \tag{13}$$

The maximum is taken over later optimal options that are strictly more efficient than $I_t(i)$, and $\max \varnothing := 0$. The two terms are the misordering loss and the per-test loss of inefficient options described in Section 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 bounds over the L options yields $\tilde{O}(L\sqrt{T})$.

5 Removing the Assumptions of CC-UCB

5.1 Removing the Cost Lower Bound: 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.

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:

$$\rho_i := \begin{cases} \dfrac{\theta_i}{c_i}, & \text{if } c_i > 0, \\ +\infty, & \text{if } c_i = 0 \wedge \theta_i > 0. \end{cases} \tag{14}$$

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 $\hat{c}_{i,t}$, initializes them from the initial observations, and uses $u_{i,t}$ and $U_{i,t} = \hat{\theta}_{i,t} + u_{i,t}$. Unlike CC-UCB, the lower confidence bound on the cost is not truncated by a positive constant:

$$L_{i,t} = \max\{\hat{c}_{i,t} - u_{i,t}, 0\}. \tag{15}$$

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 $L_{i,t} = 0$ and $U_{i,t} > 0$, or $L_{i,t} > 0$ and $U_{i,t}/L_{i,t} \ge 1$. It then orders the group $I_t^0 := \{i \in I_t : L_{i,t} = 0\}$, whose optimistic efficiency ratio is undefined, in descending order of $U_{i,t}$, and the group $I_t^+ := \{i \in I_t : L_{i,t} > 0\}$ in descending order of $U_{i,t}/L_{i,t}$, and tests the concatenated list $I_t = I_t^0 \| I_t^+$ until 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.

Theorem 3. *For sufficiently large T, the regret of CC-UCBv2 satisfies*

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

Proof sketch. The detailed proof is given in Appendix D. With the extended-real efficiency, the per-round decomposition of CC-UCB still holds. If a zero-cost option with positive success probability is placed after a positive-cost option i, the efficiency-gap factor equals one, so the misordering loss per test of i is c_i. Lemma 2 shows that this can happen only while $N_{i,t} \le 6\log t/c_i^2$. This case therefore contributes a term of the form $\min\{O(\log T/\Delta^2), T\}\Delta$ with $\Delta = c_i$, as in the argument for Theorem 2. Options with infinite efficiency incur neither misordering nor inefficient-option losses. Bounding each term by $O(\sqrt{T\log T})$ and summing over the options gives $\tilde{O}(L\sqrt{T})$. If also $\theta_i < 1$ for all i, CC-UCBv2 attains the bound of Theorem 1 with the same constant (Appendix D).

5.2 Removing the Assumption $\theta_i < 1$: BC-CC-UCBv2

We now also remove the assumption $\theta_i < 1$. Once a sure-success option is tested, the round stops, so the optimal list ends at the cheapest sure-success option. Let

$$c^* := \min\big(\{1\} \cup \{c_i : \theta_i = 1\}\big), \qquad \Delta_i^* := c_i - c^*\theta_i. \tag{17}$$

Inserting option i immediately before a suffix that always succeeds with expected cost c^* changes the expected reward by $\theta_i c^* - c_i = -\Delta_i^*$, so Δ_i^* is the loss of testing i relative to stopping at the boundary; we call it the boundary gap of option i. If $\theta_i < 1$ for all i, then $c^* = 1$, the boundary is the efficiency threshold one, and $\Delta_i^* = c_i - \theta_i$.

BC-CC-UCBv2 (Algorithm 3) modifies CC-UCBv2 so that every option that has not yet failed is treated as a possible sure-success option. Let $F_{i,t}$ be the number of failures of option i observed before round t, and let $\mathcal{Z}_t := \{i : F_{i,t} = 0\}$ and $\mathcal{A}_t := [L] \setminus \mathcal{Z}_t$. In addition to $U_{i,t}$ and $L_{i,t} = \max\{\hat{c}_{i,t} - u_{i,t}, 0\}$ of CC-UCBv2, the algorithm uses the upper confidence bound $V_{i,t} := \min\{\hat{c}_{i,t} + u_{i,t}, 1\}$ on the cost, and an auxiliary option 0 with $L_{0,t} = V_{0,t} := 1$, which represents stopping. It selects the terminal candidate $b_t \in \arg\min_{i\in\mathcal{Z}_t\cup\{0\}} L_{i,t}$, bounds the boundary cost by $V_{*,t} := \min_{i\in\mathcal{Z}_t\cup\{0\}} V_{i,t}$, and forms the prefix

$$P_t := \{i \in \mathcal{A}_t : L_{i,t} \le U_{i,t}V_{*,t}\}. \tag{18}$$

The prefix is ordered by the rule of CC-UCBv2, and b_t is appended at the end unless $b_t = 0$. These rules change CC-UCBv2 in two ways. Among the options that have never failed, at most one, b_t, is included in the list, and it is placed at the end. The inclusion condition of CC-UCBv2 is equivalent to $L_{i,t} \le U_{i,t}$ because $U_{i,t} > 0$, and BC-CC-UCBv2 replaces it by $L_{i,t} \le U_{i,t}V_{*,t}$.

Theorem 4. *For every instance with $\theta_i, c_i \in [0,1]$, the regret of BC-CC-UCBv2 satisfies*

$$R_T \le \sum_{i\in[L]:\Delta_i^*>0} \frac{192\log T}{\Delta_i^*} + O(1). \tag{19}$$

In particular, if $\theta_i < 1$ for all i, then $R_T \le \sum_{i\in[L]:\rho_i<1} \frac{192\log T}{c_i-\theta_i} + O(1)$. Here, the $O(1)$ term may depend on the problem instance but not on the horizon T.

Proof sketch. The detailed proof is given in Appendix E. Every option in the prefix has failed at least once, so it is reached with probability at least $q := \prod_{i:\theta_i<1}(1-\theta_i) > 0$. The bound $V_{*,t}$ on c^* can err in two directions. It can be too low while an option with $\theta_i < 1$ and $c_i < c^*$ has not yet failed. The first round at which every such option has failed at least once has finite expectation. From that round on, $V_{*,t} \ge c^*$ on the good event. It can also be too high while the cheapest sure-success options have few observations. These options are tested linearly often, so $V_{*,t} - c^* = O(\sqrt{\log t/t})$. Rounds on which either guarantee fails contribute $O(1)$. On the remaining rounds, the decomposition of Section 4 splits the regret into three parts. Misordering among options with $\Delta_i^* \le 0$ contributes $O(1)$. Including an option with $\Delta_i^* > 0$ contributes $192\log T/\Delta_i^*$, because such an option can be in P_t only while $N_{i,t} \le 96\log t/(\Delta_i^*)^2$. A costly sure-success option placed at the end contributes $6\log T/\Delta_i^*$.

6 Minimax Lower Bound

Finally, we show that CCB inherits the minimax lower bound of MAB.

Theorem 5. *For any algorithm, there exists a CCB instance such that*

$$R_T \ge \Omega(\sqrt{LT}). \tag{20}$$

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 $1 - c_i$. Although CCB also permits the empty list, which has expected reward 0, the arm means in the lower-bound construction of Auer et al. [1995] lie in a neighborhood of $1/2$. Hence the empty list is always dominated, and adding it to the action set does not change the minimax lower bound. 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. □

7 Numerical Experiments

We use two synthetic experiments to investigate the modifications to CC-UCB. The first examines how a misspecified cost lower bound distorts efficiency estimates, and how removing the lower-bound parameter avoids this distortion. The second compares the behavior of CC-UCBv2 and BC-CC-UCBv2 with the CC-UCB baseline in an instance containing zero-cost and sure-success options. Prior work also investigated system parameters and real-world click logs [Zhou et al., 2018]; here we focus on these two algorithmic questions.

For each option, successes and costs are independent Bernoulli variables with means θ_i and c_i, respectively. We use the exploration constant 1.5, horizon $T = 200{,}000$, and 30 independent runs. Within each experiment, the algorithms share potential outcomes indexed by round and option. Following Section 4.1, each option has one initial observation at $t = 0$, and regret is accumulated over decision rounds $t = 1, \ldots, T$. For each trajectory, we record $\sum_{t=1}^{T}[\mu(I^*) - \mu(I_t)]$ using the true means; its Monte Carlo average estimates R_T without realized-reward noise. Curves show the average over runs, and shaded bands show one sample standard deviation, with the displayed lower band clipped at zero. Line styles and marker shapes distinguish the algorithms independently of color. The labels A–E identify experimental options separately from the efficiency-ordered indices used in the analysis. All experiments are CPU-based simulations reproducible on a standard laptop or desktop. The code is available at `https://github.com/tamakoshi2001/neurips2026-cost-aware-cascading`. Appendix F studies a refined ordering for CC-UCBv2.

Avoiding distortion from a misspecified cost lower bound. For options A, B, and C, in label order, we set

$$\boldsymbol{\theta} = (0.4, 0.6, 0.3), \qquad \boldsymbol{c} = (0.08, 0.20, 0.40).$$

Their efficiencies are 5, 3, and 0.75, so the optimal list is (A, B). We compare CC-UCBv2 with CC-UCB using $\epsilon = 0.04$, half the minimum expected cost and a valid lower bound, and $\epsilon = 0.20$, which incorrectly uses B's cost as a lower bound for A. Under the latter choice, the limiting optimistic efficiencies of A and B are $0.4/0.20 = 2$ and $0.6/0.20 = 3$, reversing their correct order. This persistent reversal costs $\theta_A c_B - \theta_B c_A = 0.032$ in expected reward per round. Figure 1 shows the resulting approximately linear regret under the misspecified bound. CC-UCBv2 and CC-UCB with the valid bound have similar, much smaller regret. This illustrates the benefit of removing the lower-bound parameter rather than relying on its correct specification.

Behavior relative to the CC-UCB baseline. We set A's cost to zero and add two sure-success options D and E:

$$\boldsymbol{\theta} = (0.4, 0.6, 0.3, 1, 1), \qquad \boldsymbol{c} = (0, 0.20, 0.40, 0.25, 0.50).$$

D is the cheaper sure-success option, so $c^* = 0.25$ and the optimal list is (A, D). B is efficient relative to stopping without a success, but testing it before D loses $c_B - c^*\theta_B = 0.05$ in expected reward conditional on reaching B. E costs twice as much as D and provides a second terminal candidate to learn about. The non-certain success probabilities remain moderate so that later options can be observed. We compare CC-UCB, CC-UCBv2, and BC-CC-UCBv2, using $\epsilon = 10^{-5}$ for CC-UCB to keep the clipping effect small. Since A has zero cost, no positive ϵ satisfies CC-UCB's cost assumption in this instance, so CC-UCB serves only as an empirical baseline. CC-UCBv2 uses the ordering rule in Algorithm 2. Figure 2 shows similar regret for CC-UCB and CC-UCBv2. BC-CC-UCBv2 incurs a larger exploratory cost, followed by much slower regret accumulation. The larger standard deviation of BC-CC-UCBv2 mainly reflects run-to-run variation in the number of rounds for which the boundary criterion retains B in the prefix.

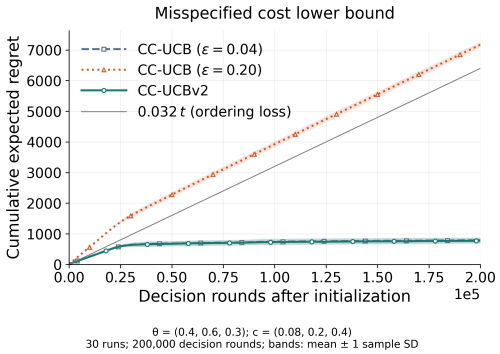

Figure 1: Effect of a misspecified cost lower bound on efficiency estimates. Curves show mean regret; bands show one standard deviation across 30 runs.

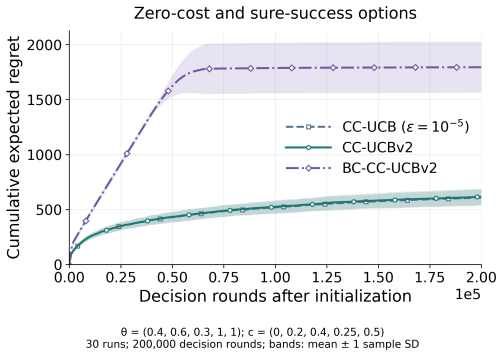

Figure 2: CC-UCB and its modifications with a zero-cost option and two sure-success options. Curves and bands use the same convention as Figure 1.

8 Limitations

Our results leave several limitations. First, the problem-independent upper bound $\tilde{O}(L\sqrt{T})$ and the lower bound $\Omega(\sqrt{LT})$ differ by a factor of order $\sqrt{L}$, up to logarithmic factors. The lower bound is inherited from the MAB subclass and does not use the cascading structure. 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, the match between our problem-dependent upper bound and the lower bound of Zhou et al. [2018] holds only when $\theta_i < 1$ for all i; when sure-success options exist, we do not provide a matching lower bound. Moreover, we do not provide a problem-independent bound for BC-CC-UCBv2, and the constant 192 in Theorem 4 is larger than the constant 72 that CC-UCB and CC-UCBv2 attain when $\theta_i < 1$ for all i.

Third, 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.

Fourth, the numerical experiments cover only a few synthetic instances and do not provide an exhaustive empirical comparison.

Finally, the broader impact of this work depends on the application in which sequential testing is deployed. Cost-aware sequential testing can reduce the time, signaling, or energy spent on tests, as in the handover example of Section 1. In high-stakes domains, deployment should validate the modeling assumptions and cost definitions, since misuse could amplify decision errors.

9 Conclusion

We developed a regret decomposition for CCB that separates the regret of testing inefficient options from that of misordering options, improving the dependence of the CC-UCB bound on the inefficiency gap $c_i - \theta_i$. CC-UCBv2 removes the known cost lower bound while preserving both bounds, and BC-CC-UCBv2 further removes the assumption $\theta_i < 1$ with a problem-dependent bound in terms of the boundary gap $c_i - c^*\theta_i$. We also showed that the minimax regret is bounded from below by $\Omega(\sqrt{LT})$.

Acknowledgments and Disclosure of Funding

SI is supported by JSPS KAKENHI Grant Number JP25K03184 and by JST PRESTO, Japan, Grant Number JPMJPR2511. The authors declare no competing interests.

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.

Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont, and Boammani Aser Lompo. Parametric graph for unimodal ranking bandit. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3630–3639. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/gauthier21a.html.

Camille-Sovanneary Gauthier, Romaric Gaudel, and Elisa Fromont. UniRank: Unimodal bandit algorithms for online ranking. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7279–7309. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/gauthier22a.html.

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 2015a. PMLR. URL https://proceedings.mlr.press/v37/kveton15.html.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/1f50893f80d6830d62765ffad7721742-Paper.pdf.

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

Tor Lattimore, Branislav Kveton, Shuai Li, and Csaba Szepesvari. Toprank: A practical algorithm for online stochastic ranking. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/de03beffeed9da5f3639a621bcab5dd4-Paper.pdf.

Chang Li, Branislav Kveton, Tor Lattimore, Ilya Markov, Maarten de Rijke, Csaba Szepesvári, and Masrour Zoghi. Bubblerank: Safe online learning to re-rank via implicit click feedback. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 196–206. PMLR, 22–25 Jul 2020a. URL https://proceedings.mlr.press/v115/li20b.html.

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 2020b.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952.

Hiroaki Shiino, Kaito Ariu, Kenshi Abe, and Riku Togashi. Exploration of unranked items in safe online learning to re-rank. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1991–1995, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591985. URL https://doi.org/10.1145/3539618.3591985.

Daniel Vial, Sujay Sanghavi, Sanjay Shakkottai, and R. Srikant. Minimax regret for cascading bandits. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 29126–29138. Curran Associates, Inc., 2022. doi: 10.52202/068431-2112. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bbb397739546de027bc81a2c2a8fb119-Paper-Conference.pdf.

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.

A Algorithms

This section presents the detailed descriptions of the algorithms analyzed in this paper: CC-UCB, CC-UCBv2, and BC-CC-UCBv2. All 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.

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

Require: $\epsilon > 0$

Initialization: for each $i \in [L]$, use the initial observation $(X_{i,0}, Y_{i,0})$ (Section 4.1) and set
    $N_{i,1} \leftarrow 1,\ \hat{\theta}_{i,1} \leftarrow X_{i,0},\ \hat{c}_{i,1} \leftarrow Y_{i,0}$.
for $t = 1, 2, \ldots$ **do**
    $I_t \leftarrow \varnothing$
    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
    $\tilde{I}_t \leftarrow \varnothing$
    for $k = 1, 2, \ldots, |I_t|$ **do**
        test option $I_t(k)$; observe $(X_{I_t(k),t}, Y_{I_t(k),t})$
        $\tilde{I}_t \leftarrow \tilde{I}_t \| (I_t(k))$
        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 (Well-Defined for Zero-Cost Options)

Initialization: for each $i \in [L]$, use the initial observation $(X_{i,0}, Y_{i,0})$ (Section 4.1) and set
    $N_{i,1} \leftarrow 1,\ \hat{\theta}_{i,1} \leftarrow X_{i,0},\ \hat{c}_{i,1} \leftarrow Y_{i,0}$.
for $t = 1, 2, \ldots$ **do**
    $I_t \leftarrow \varnothing$
    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
    Partition I_t into two disjoint sets:
        $I_t^0 \leftarrow \{i \in I_t : L_{i,t} = 0\}$ ▷ Undefined ratio group
        $I_t^+ \leftarrow \{i \in I_t : L_{i,t} > 0\}$ ▷ Well-defined ratio group
    Sort options in I_t^0 in descending order of $U_{i,t}$ ▷ By success probability
    Sort options in I_t^+ in descending order of $\dfrac{U_{i,t}}{L_{i,t}}$ ▷ By efficiency ratio
    Re-define I_t as the concatenation:
        $I_t \leftarrow I_t^0 \| I_t^+$ ▷ Hierarchical ordering
    $\tilde{I}_t \leftarrow \varnothing$
    for $k = 1, 2, \ldots, |I_t|$ **do**
        test option $I_t(k)$; observe $(X_{I_t(k),t}, Y_{I_t(k),t})$
        $\tilde{I}_t \leftarrow \tilde{I}_t \| (I_t(k))$
        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

Algorithm 3 Boundary-Certifying CC-UCBv2 (BC-CC-UCBv2)

Initialization: for each $i \in [L]$, use the initial observation $(X_{i,0}, Y_{i,0})$ (Section 4.1) and set
    $N_{i,1} \leftarrow 1,\ \hat{\theta}_{i,1} \leftarrow X_{i,0},\ \hat{c}_{i,1} \leftarrow Y_{i,0},\ F_{i,1} \leftarrow 1 - X_{i,0}.$
for $t = 1, 2, \ldots$ **do**
    for $i = 1, 2, \ldots, L$ **do**
        $u_{i,t} \leftarrow \sqrt{\dfrac{1.5 \log t}{N_{i,t}}}, \quad U_{i,t} \leftarrow \hat{\theta}_{i,t} + u_{i,t}$
        $L_{i,t} \leftarrow \max\{\hat{c}_{i,t} - u_{i,t},\ 0\}, \quad V_{i,t} \leftarrow \min\{\hat{c}_{i,t} + u_{i,t},\ 1\}$ ▷ Cost LCB and UCB
    end for
    $\mathcal{Z}_t \leftarrow \{i \in [L] : F_{i,t} = 0\}, \quad \mathcal{A}_t \leftarrow [L] \setminus \mathcal{Z}_t, \quad L_{0,t} \leftarrow 1, \quad V_{0,t} \leftarrow 1$
    $b_t \leftarrow \arg\min_{i \in \mathcal{Z}_t \cup \{0\}} L_{i,t}$ ▷ Ties: actual options first, then smaller $N_{i,t}$
    $V_{*,t} \leftarrow \min_{i \in \mathcal{Z}_t \cup \{0\}} V_{i,t}$ ▷ Confidence bound on c^*
    $P_t \leftarrow \{i \in \mathcal{A}_t : L_{i,t} \le U_{i,t} V_{*,t}\}$ ▷ Prefix candidates
    $P_t^0 \leftarrow \{i \in P_t : L_{i,t} = 0\}$ sorted in descending order of $U_{i,t}$
    $P_t^+ \leftarrow \{i \in P_t : L_{i,t} > 0\}$ sorted in descending order of $U_{i,t}/L_{i,t}$
    $I_t \leftarrow P_t^0 \| P_t^+; \quad$ **if** $b_t \neq 0$ **then** $I_t \leftarrow I_t \| (b_t)$
    $\tilde{I}_t \leftarrow \varnothing$
    for $k = 1, 2, \ldots, |I_t|$ **do**
        test option $I_t(k)$; observe $(X_{I_t(k),t}, Y_{I_t(k),t})$
        $\tilde{I}_t \leftarrow \tilde{I}_t \| (I_t(k))$
        if $X_{I_t(k),t} = 1$ **then**
            break
        end if
    end for
    for all $i \in \tilde{I}_t$ **do**
        $N_{i,t+1} \leftarrow N_{i,t} + 1, \quad F_{i,t+1} \leftarrow F_{i,t} + 1 - X_{i,t}$
        $\hat{\theta}_{i,t+1} \leftarrow \dfrac{N_{i,t}\hat{\theta}_{i,t} + X_{i,t}}{N_{i,t+1}}, \quad \hat{c}_{i,t+1} \leftarrow \dfrac{N_{i,t}\hat{c}_{i,t} + Y_{i,t}}{N_{i,t+1}}$
    end for
    for all $i \notin \tilde{I}_t$ **do**
        $N_{i,t+1} \leftarrow N_{i,t},\ F_{i,t+1} \leftarrow F_{i,t},\ \hat{\theta}_{i,t+1} \leftarrow \hat{\theta}_{i,t},\ \hat{c}_{i,t+1} \leftarrow \hat{c}_{i,t}$
    end for
end for

B Auxiliary Lemmas

The following facts are shared by the analyses of all three algorithms. Every list I satisfies $-L \le \mu(I) \le 1$, so $\mathrm{reg}_t \le L+1$ for every t. The concentration bound of Lemma 1 below is informative only for $t \ge 2$, because $\log 1 = 0$. We therefore bound the regret of the first round by $L+1$, which gives

$$R_T \le L + 1 + \sum_{t=2}^{T} \mathbb{E}[\mathrm{reg}_t]. \tag{21}$$

Each algorithm starts from $N_{i,1} = 1$, so $N_{i,t} \in [t]$ for every t. As in the standard analysis of UCB algorithms, we may assume that, for each option i, the initial observation and the outcomes of its successive tests are the successive entries of an i.i.d. sequence $((X_i^{(n)}, Y_i^{(n)}))_{n\ge1}$ drawn in advance, independently across options; this does not change the law of the process.

Lemma 1 (Error rounds)**.** *For each of the three algorithms, $B_\infty := \sum_{t\ge2} \mathbf{1}\{\mathcal{E}_t\}$ satisfies $\mathbb{E}[B_\infty] \le 4L\sum_{t\ge2} t^{-2} < 4L$.*

Proof. Fix i and $t \ge 2$. For each $n \in [t]$, Hoeffding's inequality applied to the first n entries of the i.i.d. sequence of option i shows that their empirical success rate deviates from θ_i by more than $\sqrt{1.5\log t/n}$ with probability at most $2\exp(-3\log t) = 2t^{-3}$. Since $N_{i,t} \in [t]$, a union bound over $n \in [t]$ gives $\mathbb{P}(|\hat\theta_{i,t} - \theta_i| > u_{i,t}) \le 2t^{-2}$. The same holds for the cost, so $\mathbb{P}(\mathcal{E}_t) \le 4Lt^{-2}$. □

Lemma 2 (Threshold for misordering)**.** *Let i be an option with $c_i > 0$. For an efficiency $\lambda \in (\rho_i, +\infty]$, define*

$$g_i(\lambda) := c_i\Big(1 - \frac{\rho_i}{\lambda}\Big), \qquad n_{i,t}(\lambda) := \frac{6(1+\lambda)^2\log t}{(\lambda-\rho_i)^2 c_i^2}, \tag{22}$$

where, for $\lambda = +\infty$, these are understood as their limits $g_i(+\infty) = c_i$ and $n_{i,t}(+\infty) = 6\log t/c_i^2$. If the event $\mathcal{E}_t^c$ holds at round t and an option j with $\rho_j > \rho_i$ is placed after i under the ordering rule of CC-UCB, or under that of CC-UCBv2 when zero-cost options are allowed, then $N_{i,t} \le n_{i,t}(\rho_j)$. Here $g_i(\lambda)$ is the loss factor of placing i before an option of efficiency λ, and the claim says that this misordering can occur only while $N_{i,t} \le n_{i,t}(\lambda)$. Moreover, g_i is nondecreasing and $n_{i,t}$ is nonincreasing on $(\rho_i, +\infty]$, and

$$n_{i,t}(\lambda) \le \frac{24\log t}{g_i(\lambda)^2} \qquad \textit{whenever } \lambda \ge 1. \tag{23}$$

Proof. First suppose that $c_j > 0$. Under the CC-UCBv2 ordering rule, consider the case $L_{i,t} = 0$, in which option i belongs to I_t^0 and its position is not determined by the ratio $U_{i,t}/L_{i,t}$. On $\mathcal{E}_t^c$, we have $L_{i,t} \ge c_i - 2u_{i,t}$, so $L_{i,t} = 0$ implies $c_i - 2u_{i,t} \le 0$, that is,

$$N_{i,t} \le \frac{6\log t}{c_i^2} \le \frac{6(1+\rho_j)^2\log t}{(\rho_j-\rho_i)^2 c_i^2}, \tag{24}$$

where the last inequality follows from $0 < \rho_j - \rho_i < 1 + \rho_j$. The right-hand side is $n_{i,t}(\rho_j)$, so the claim holds in this case. In the remaining cases, the options i and j are ordered by the ratio $U_{\cdot,t}/L_{\cdot,t}$. Under CC-UCB, this holds because $L_{\cdot,t} \ge \epsilon > 0$ always. Under CC-UCBv2 with $L_{i,t} > 0$, option i belongs to I_t^+, and so does every option placed after i, including j. Since the event $\mathcal{E}_t^c$ holds, if option j is placed after i, then

$$\mathbf{1}\left\{\frac{U_{i,t}}{L_{i,t}} \ge \frac{U_{j,t}}{L_{j,t}}\right\} \le \mathbf{1}\left\{\frac{U_{i,t}}{L_{i,t}} \ge \frac{\theta_j}{c_j}\right\} \tag{25}$$

$$\le \mathbf{1}\left\{c_i - 2u_{i,t} \le 0 \ \cup\ \frac{\theta_i + 2u_{i,t}}{c_i - 2u_{i,t}} \ge \frac{\theta_j}{c_j}\right\} \tag{26}$$

$$= \mathbf{1}\left\{N_{i,t} \le \max\left\{\frac{6\log t}{c_i^2}, \frac{6(1+\rho_j)^2\log t}{(\rho_j-\rho_i)^2c_i^2}\right\}\right\} \tag{27}$$

$$= \mathbf{1}\left\{N_{i,t} \le \frac{6(1+\rho_j)^2\log t}{(\rho_j-\rho_i)^2c_i^2}\right\}. \tag{28}$$

Thus $N_{i,t} \le n_{i,t}(\rho_j)$ when $c_j > 0$.

Now suppose that option j has zero expected cost. On $\mathcal{E}_t^c$, we have $L_{j,t} \le c_j = 0$, so under the CC-UCBv2 ordering rule j belongs to I_t^0 and is placed before every option of I_t^+. Therefore, if j is placed after i, then $i \in I_t^0$, that is, $L_{i,t} = 0$. On $\mathcal{E}_t^c$,

$$\mathbf{1}\{L_{i,t} = 0\} \le \mathbf{1}\{c_i - 2u_{i,t} \le 0\} \tag{29}$$

$$= \mathbf{1}\left\{N_{i,t} \le \frac{6\log t}{c_i^2}\right\}. \tag{30}$$

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

Finally, $1 - \rho_i/\lambda$ is nondecreasing in λ, and $(1+\lambda)/(\lambda - \rho_i)$ is decreasing in $\lambda > \rho_i$ because its derivative is $-(1+\rho_i)/(\lambda-\rho_i)^2$; the limits at $+\infty$ preserve both monotonicities. For (23), use $(\lambda - \rho_i)c_i = \lambda g_i(\lambda)$ to write $n_{i,t}(\lambda) = 6(1+\lambda)^2 \log t/(\lambda^2 g_i(\lambda)^2)$ and use $1+\lambda \le 2\lambda$; for $\lambda = +\infty$, $n_{i,t}(+\infty) = 6\log t/g_i(+\infty)^2$. □

Lemma 3 (Selection of an inefficient option)**.** *If the event $\mathcal{E}_t^c$ holds at round t, then for any inefficient option i, if*

$$N_{i,t} > \frac{24\log t}{(c_i - \theta_i)^2}, \tag{31}$$

it follows that $i \notin I_t$.

Proof. If i is included in I_t, then $U_{i,t} \ge L_{i,t}$. On $\mathcal{E}_t^c$, this gives

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

$$\le \mathbf{1}\left\{N_{i,t} \le \frac{24\log t}{(c_i - \theta_i)^2}\right\}. \tag{32}$$

□

Lemma 4 (Lemma 3 of Kveton et al. [2014])**.** *Let $x_1 \ge \cdots \ge x_m > 0$. Then,*

$$\sum_{k=1}^{m-1} \frac{x_k - x_{k+1}}{x_k^2} + \frac{1}{x_m} \le \frac{2}{x_m}. \tag{33}$$

Proof. Since $x_k \ge x_{k+1}$ for all $k < m$,

$$\sum_{k=1}^{m-1} \frac{x_k - x_{k+1}}{x_k^2} + \frac{1}{x_m} \le \sum_{k=1}^{m-1} \frac{x_k - x_{k+1}}{x_k x_{k+1}} + \frac{1}{x_m} = \sum_{k=1}^{m-1}\left(\frac{1}{x_{k+1}} - \frac{1}{x_k}\right) + \frac{1}{x_m} = \frac{2}{x_m} - \frac{1}{x_1} < \frac{2}{x_m}. \tag{34}$$

□

Lemma 5. *Let*

$$x_1 \ge x_2 \ge \cdots \ge x_m > 0, \tag{35}$$

and let $a, h > 0$. Define

$$\nu_i := \begin{cases} 0 & \text{if } i = 0, \\ \min\left\{\frac{a}{x_i^2},\, h\right\} & \text{if } i \ne 0. \end{cases} \tag{36}$$

Then,

$$\sum_{i=1}^{m-1}(x_i - x_{i+1})\nu_i + x_m\nu_m = \sum_{i=1}^{m} x_i(\nu_i - \nu_{i-1}) \le 2\sqrt{ah}. \tag{37}$$

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

$$0 = \nu_0 \le \nu_1 \le \cdots \le \nu_m \le h. \tag{38}$$

We bound the left-hand side as

$$\sum_{i=1}^{m} x_i \left(\nu_i - \nu_{i-1}\right) \le \sum_{i=1}^{m} \sqrt{a} \frac{\nu_i - \nu_{i-1}}{\sqrt{\nu_i}}, \tag{39}$$

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

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

$$(\nu_i - \nu_{i-1}) \frac{1}{\sqrt{\nu_i}} \le \int_{\nu_{i-1}}^{\nu_i} \frac{1}{\sqrt{\nu}} \, d\nu. \tag{40}$$

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

$$\sum_{i=1}^{m} \sqrt{a} \frac{\nu_i - \nu_{i-1}}{\sqrt{\nu_i}} \le \sqrt{a} \int_0^h \frac{1}{\sqrt{\nu}} \, d\nu = 2\sqrt{ah}. \tag{41}$$

□

C Proof of Theorems 1 and 2

Lemma 6 (Good-event per-round regret)**.** *Let g_i and $n_{i,t}$ be as in Lemma 2. Under the assumption of Section 4, every cost is positive, so all efficiencies are finite. On the good event $\mathcal{E}_t^c$,*

$$\begin{aligned} \mathrm{reg}_t \le & \sum_{i=1}^{L} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}] \max_{\substack{j \in [L^*]:\ \rho_j > \rho_i, \\ N_{i,t} \le n_{i,t}(\rho_j)}} g_i(\rho_j) \\ & + \sum_{i \in [L] \setminus [L^*]} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}](c_i - \theta_i), \end{aligned} \tag{42}$$

where $\max \varnothing := 0$.

Proof. We first define the intermediate policies used in the telescoping argument. For a set S of options, let $\mathrm{ord}(S)$ be the list of the options of S in increasing order of index, that is, in descending order of efficiency with ties broken by the fixed indexing; Appendix E uses the same notation. For each $i = 1, \ldots, |I_t| + 1$, define

$$I_t^{(i)} := (I_t(1), \ldots, I_t(i-1)) \| \, \mathrm{ord}\big(\{I_t(i), \ldots, I_t(|I_t|)\} \cap [L^*]\big), \tag{43}$$

which keeps the first $i - 1$ options of I_t, removes the inefficient options from the remaining suffix, and sorts the rest. For $i = |I_t| + 1$ the suffix is empty, so $I_t^{(|I_t|+1)} = I_t$. On the good event $\mathcal{E}_t^c$, all optimal options are included in I_t, so $I_t^{(1)}$ consists of all optimal options in descending order of efficiency, that is, $I_t^{(1)} = I^*$.

Fix $i \in [|I_t|]$ and write $J := I_t^{(i+1)}$. The lists J and $I_t^{(i)}$ 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 the testing event $\mathcal{T}_{I_t(i),t}$. Moreover, $J(i) = I_t(i)$, and the options after position i in J are sorted. If $I_t(i) \in [L^*]$, then $I_t^{(i)}$ is obtained from J by moving $I_t(i)$ from position i to its position in the sorted order, which we denote by m. If $I_t(i) \in [L] \setminus [L^*]$, then $I_t^{(i)}$ is obtained from J by moving $I_t(i)$ to the last position $m := |J|$ and then deleting it. In both cases, the move is a sequence of adjacent swaps, and for $s = 1, \ldots, m - i$ the s-th swap exchanges $I_t(i)$ and $J(i + s)$. Since position i is reached with probability $\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}]$, (5) shows that this swap changes the expected reward by

$$\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] \prod_{l=1}^{s-1} (1 - \theta_{J(i+l)})(c_{I_t(i)} \theta_{J(i+s)} - c_{J(i+s)} \theta_{I_t(i)}). \tag{44}$$

The deletion, which is performed only if $I_t(i) \in [L] \setminus [L^*]$, then increases the expected reward by

$$\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] \prod_{l=i+1}^{m} (1 - \theta_{I_t^{(i+1)}(l)})(c_{I_t(i)} - \theta_{I_t(i)}) \mathbf{1}\{I_t(i) \in [L] \setminus [L^*]\}. \tag{45}$$

Indeed, after reaching the deleted option, keeping it would contribute the single-option expected reward $\theta_{I_t(i)} - c_{I_t(i)}$, and the product is the probability that all options placed before it after the swaps fail. Summing the changes from the swaps and bounding the product in the deletion term by one, we obtain

$$\begin{aligned}
&\mu(I_t^{(i)}) - \mu(I_t^{(i+1)}) \\
&\le \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] \sum_{k=i+1}^{m} \prod_{l=i+1}^{k-1} (1 - \theta_{I_t^{(i+1)}(l)})(\theta_{I_t^{(i+1)}(k)} c_{I_t(i)} - c_{I_t^{(i+1)}(k)} \theta_{I_t(i)}) \\
&\quad + \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}) \mathbf{1}\{I_t(i) \in [L] \setminus [L^*]\}. \qquad (46)
\end{aligned}$$

For the swap term, use

$$\theta_{I_t^{(i+1)}(k)} c_{I_t(i)} - c_{I_t^{(i+1)}(k)} \theta_{I_t(i)} = \theta_{I_t^{(i+1)}(k)} c_{I_t(i)} \left(1 - \frac{\rho_{I_t(i)}}{\rho_{I_t^{(i+1)}(k)}}\right), \qquad (47)$$

which is valid whenever $\rho_{I_t^{(i+1)}(k)} > \rho_{I_t(i)}$. If instead $\rho_{I_t^{(i+1)}(k)} \le \rho_{I_t(i)}$, the swap term is nonpositive and we drop it. Every option passed by $I_t(i)$ belongs to $\{I_t(i') : i' > i\} \cap [L^*]$. Since

$$\sum_{k=i+1}^{m} \prod_{l=i+1}^{k-1} (1 - \theta_{I_t^{(i+1)}(l)}) \theta_{I_t^{(i+1)}(k)} \le 1,$$

we obtain, for the right-hand side (RHS) of (46),

$$\begin{aligned}
\text{RHS of (46)} &\le \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] c_{I_t(i)} \max_{\substack{k \in \{I_t(i'):i'>i\} \cap [L^*] \\ \rho_k > \rho_{I_t(i)}}} \left(1 - \frac{\rho_{I_t(i)}}{\rho_k}\right) \\
&\quad + \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}) \mathbf{1}\{I_t(i) \in [L] \setminus [L^*]\}. \qquad (48)
\end{aligned}$$

Here $\max \varnothing := 0$, which covers the case $m = i$ and the case where no passed option is strictly more efficient than $I_t(i)$.

Every option k in the maximum of (48) belongs to $[L^*]$, satisfies $\rho_k > \rho_{I_t(i)}$, and is placed after $I_t(i)$, so Lemma 2 gives $N_{I_t(i),t} \le n_{I_t(i),t}(\rho_k)$. Since $c_{I_t(i)}(1 - \rho_{I_t(i)}/\rho_k) = g_{I_t(i)}(\rho_k)$, the maximum is at most the maximum of $g_{I_t(i)}(\rho_j)$ over $j \in [L^*]$ with $\rho_j > \rho_{I_t(i)}$ and $N_{I_t(i),t} \le n_{I_t(i),t}(\rho_j)$. A telescoping sum over the suffix-sorting sequence therefore yields

$$\begin{aligned}
\mathrm{reg}_t \le &\sum_{i=1}^{|I_t|-1} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}] \max_{\substack{j \in [L^*]:\ \rho_j > \rho_{I_t(i)}, \\ N_{I_t(i),t} \le n_{I_t(i),t}(\rho_j)}} g_{I_t(i)}(\rho_j) \\
&+ \sum_{I_t(i) \in [L] \setminus [L^*]} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(i),t}\}](c_{I_t(i)} - \theta_{I_t(i)}). \qquad (49)
\end{aligned}$$

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 (42). □

Lemma 7 (Regret on error rounds)**.** *For CC-UCB and CC-UCBv2,*

$$\mathbb{E}\left[\sum_{t=2}^{T} \mathrm{reg}_t \mathbf{1}\{\mathcal{E}_t\}\right] \le 4L(L+1). \qquad (50)$$

Proof. Since $\mathrm{reg}_t \le L+1$, the left-hand side is at most $(L+1)\mathbb{E}[B_\infty]$, and Lemma 1 gives (50). □

The next lemma bounds how often an optimal option has few observations on good rounds. Lemma 3 of Zhou et al. [2018] proves such a bound for the threshold associated with the preceding optimal option. We follow its argument for a threshold $A \log t$ with an arbitrary constant A.

Lemma 8 (Few observations of an optimal option)**.** *Consider CC-UCB under the assumption of Section 4, and assume $\theta_k < 1$ for all $k \in [L]$. Fix $i \in [L^*]$, let $p_i := \prod_{k\in[L]\setminus\{i\}}(1-\theta_k) > 0$, and let $A > 0$. Define*

$$\zeta := \max\left\{n \in \mathbb{N} : A(\log n + \log 2) \geq \frac{p_i}{2}n\right\},$$

with $\max \varnothing := 0$. Then, for every T,

$$\mathbb{E}\left[\sum_{t=2}^{T} \mathbf{1}\{\mathcal{E}_t^c\}\mathbf{1}\{N_{i,t} \leq A\log t\}\right] \leq \zeta + \frac{2}{p_i^2} + 4L. \tag{51}$$

Proof. The argument uses the outcomes of untested options at a given round, which the i.i.d. sequences of Appendix B do not provide. Since the claim concerns only the law of the process, we may instead use the following representation: at every round t, the outcomes and costs $(X_{k,t}, Y_{k,t})_{k\in[L]}$ are drawn for all options, independently across options and rounds, and only those of the tested options are observed. Let $\widetilde{\mathcal{F}}_t$ be the σ-algebra generated by all outcomes and costs drawn before round t, including unobserved ones. Then $\mathcal{F}_t \subseteq \widetilde{\mathcal{F}}_t$, and $(X_{k,t})_{k\in[L]}$ is independent of $\widetilde{\mathcal{F}}_t$.

Let $\tau_1 < \tau_2 < \cdots$ be the rounds $t \geq 2$ at which $\mathcal{E}_t^c$ occurs. Since $\mathcal{E}_t \in \mathcal{F}_t \subseteq \widetilde{\mathcal{F}}_t$, each τ_n is a stopping time with respect to $(\widetilde{\mathcal{F}}_t)_t$, and $\tau_n < \infty$ almost surely because $\mathbb{E}[B_\infty] < \infty$ by Lemma 1. Let $Z_t := \mathbf{1}\{X_{k,t} = 0\ \forall k \in [L] \setminus \{i\}\}$ and $S_n := \sum_{m=1}^{n} Z_{\tau_m}$.

Lower bound on $N_{i,t}$. On $\mathcal{E}_t^c$, we have $U_{k,t} \geq \theta_k$ and $L_{k,t} \leq c_k$ for every k, since $c_k > \epsilon$. Hence $U_{i,t}/L_{i,t} \geq \rho_i \geq 1$, and $i \in I_t$. If moreover $Z_t = 1$, then every option placed before i in I_t fails, so $\mathcal{T}_{i,t}$ occurs and i is tested at round t. Since $N_{i,1} = 1$ and $Z_{\tau_n} \leq 1$,

$$N_{i,\tau_n} \geq 1 + \sum_{m=1}^{n-1} Z_{\tau_m} \geq S_n. \tag{52}$$

Concentration of S_n. The variable Z_t is $\widetilde{\mathcal{F}}_{t+1}$-measurable and independent of $\widetilde{\mathcal{F}}_t$, with $\mathbb{P}(Z_t = 1) = p_i$. Since τ_m is a stopping time, we have $\mathbb{E}[Z_{\tau_m} \mid \widetilde{\mathcal{F}}_{\tau_m}] = p_i$. As $\tau_{m-1} + 1 \leq \tau_m$, the variables $Z_{\tau_m} - p_i$ form a martingale difference sequence with respect to $(\widetilde{\mathcal{F}}_{\tau_m+1})_{m\geq 1}$, and each takes values in an interval of length one. The Azuma–Hoeffding inequality gives

$$\mathbb{P}\Big(S_n \leq \frac{p_i}{2}n\Big) \leq \exp\Big(-\frac{p_i^2}{2}n\Big). \tag{53}$$

Counting. Each round counted on the left-hand side below is τ_n for some n with $N_{i,\tau_n} \leq A\log\tau_n$. We show that such an n satisfies $S_n \leq A(\log n + \log 2)$ or $B_\infty \geq n$. If $\tau_n \leq 2n$, then (52) gives $S_n \leq N_{i,\tau_n} \leq A\log\tau_n \leq A(\log n + \log 2)$. If $\tau_n > 2n$, then at least $\tau_n - 1 - n \geq n$ of the rounds $2, \ldots, \tau_n$ are error rounds, so $B_\infty \geq n$. Hence

$$\sum_{t=2}^{T} \mathbf{1}\{\mathcal{E}_t^c\}\mathbf{1}\{N_{i,t} \leq A\log t\} \leq \sum_{n\geq 1} \mathbf{1}\{S_n \leq A(\log n + \log 2)\} + \sum_{n\geq 1} \mathbf{1}\{B_\infty \geq n\}.$$

The last sum equals B_∞, whose expectation is less than $4L$ by Lemma 1. For $n > \zeta$, we have $A(\log n + \log 2) < p_i n/2$, so (53) bounds the expectation of the first sum by

$$\zeta + \sum_{n\geq 1} \exp\Big(-\frac{p_i^2}{2}n\Big) \leq \zeta + \frac{2}{p_i^2},$$

where we used $\sum_{n\geq 1} e^{-an} = 1/(e^a - 1) \leq 1/a$ for $a > 0$. □

Proof of Theorems 1 and 2. We first isolate the regret on the error events. By (21),

$$R_T \leq L + 1 + \mathbb{E}\left[\sum_{t=2}^{T} \mathrm{reg}_t \mathbf{1}\{\mathcal{E}_t^c\}\right] + \mathbb{E}\left[\sum_{t=2}^{T} \mathrm{reg}_t \mathbf{1}\{\mathcal{E}_t\}\right]. \tag{54}$$

Applying Lemmas 6 and 7 to the two terms gives

$$\text{RHS of (54)} \le \mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{L}\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\max_{\substack{j\in[L^*]:\ \rho_j>\rho_i,\\ N_{i,t}\le n_{i,t}(\rho_j)}} g_i(\rho_j)\,\mathbf{1}\{\mathcal{E}_t^c\}\right] + \mathbb{E}\left[\sum_{t=2}^{T}\sum_{i\in[L]\setminus[L^*]}\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}](c_i-\theta_i)\mathbf{1}\{\mathcal{E}_t^c\}\right] + (L+1)(4L+1). \tag{55}$$

It remains to rewrite the first term in a form that separates adjacent efficiency gaps. For each option i, let $s_i := |\{j \in [L^*] : \rho_j > \rho_i\}|$. Since the options are indexed in descending order of efficiency, this set is $[s_i]$. If $s_i \ge 1$, then $\rho_1 \ge \cdots \ge \rho_{s_i} > \rho_i$, so Lemma 2 gives $g_i(\rho_1) \ge \cdots \ge g_i(\rho_{s_i}) > 0$ and $n_{i,t}(\rho_1) \le \cdots \le n_{i,t}(\rho_{s_i})$. We claim that

$$\max_{\substack{j\in[s_i]:\\ N_{i,t}\le n_{i,t}(\rho_j)}} g_i(\rho_j) \le \sum_{j=1}^{s_i-1}\big(g_i(\rho_j)-g_i(\rho_{j+1})\big)\mathbf{1}\{N_{i,t}\le n_{i,t}(\rho_j)\} + g_i(\rho_{s_i})\mathbf{1}\{N_{i,t}\le n_{i,t}(\rho_{s_i})\}. \tag{56}$$

If no $j \in [s_i]$ satisfies $N_{i,t} \le n_{i,t}(\rho_j)$, the left-hand side is zero and every term on the right is nonnegative. Otherwise, let j° be the smallest such j; the maximum equals $g_i(\rho_{j^\circ})$ because $g_i(\rho_j)$ is nonincreasing in j. Since $n_{i,t}(\rho_j)$ is nondecreasing in j, $N_{i,t} \le n_{i,t}(\rho_j)$ holds for every $j \in \{j^\circ, \dots, s_i\}$, and the right-hand side is at least $\sum_{j=j^\circ}^{s_i-1}(g_i(\rho_j) - g_i(\rho_{j+1})) + g_i(\rho_{s_i}) = g_i(\rho_{j^\circ})$. Define the expected number of tests

$$K_{i,j} := \mathbb{E}\left[\sum_{t=2}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{N_{i,t}\le n_{i,t}(\rho_j)\}\mathbf{1}\{\mathcal{E}_t^c\}\right] \le T.$$

Substituting (56) and using the tower property yields the following bound, which we use for both theorems:

$$R_T \le \sum_{i=1}^{L} M_i + \sum_{i\in[L]\setminus[L^*]}(c_i-\theta_i)\,\mathbb{E}\left[\sum_{t=2}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{\mathcal{E}_t^c\}\right] + (L+1)(4L+1), \tag{57}$$

where

$$M_i := \sum_{j=1}^{s_i-1}\big(g_i(\rho_j)-g_i(\rho_{j+1})\big)K_{i,j} + g_i(\rho_{s_i})K_{i,s_i} \tag{58}$$

is the misordering term of option i, and $M_i := 0$ if $s_i = 0$.

Every $j \in [s_i]$ satisfies $j \le L^*$, hence $\rho_j \ge 1$, and (23) gives $n_{i,T}(\rho_j) \le 24\log T/g_i(\rho_j)^2$ for $j \in [s_i]$. Moreover, $K_{i,j} \le n_{i,T}(\rho_j)$, because $n_{i,t}(\rho_j) \le n_{i,T}(\rho_j)$, $N_{i,t}$ increases by one whenever option i is tested, and option i is tested at most once at each round. We now derive the two regret bounds from (57).

Case 1: Problem-dependent bound We bound the three contributions in (57): misordering inefficient options, selecting inefficient options, and misordering optimal options.

Misordering inefficient options. Fix $i \in [L] \setminus [L^*]$ with $s_i \ge 1$. By the bounds on $K_{i,j}$ above and Lemma 4 with $x_j = g_i(\rho_j)$,

$$M_i \le 24\log T\left(\sum_{j=1}^{s_i-1}\frac{g_i(\rho_j)-g_i(\rho_{j+1})}{g_i(\rho_j)^2} + \frac{1}{g_i(\rho_{s_i})}\right) \le \frac{48\log T}{g_i(\rho_{s_i})} \le \frac{48\log T}{c_i-\theta_i}, \tag{59}$$

where the last inequality uses $\rho_{s_i} \ge 1$, so that $g_i(\rho_{s_i}) \ge c_i(1-\rho_i) = c_i - \theta_i$.

Selecting inefficient options. By Lemma 3, an inefficient option i is tested on a round with $\mathcal{E}_t^c$ only while $N_{i,t} \le 24\log T/(c_i-\theta_i)^2$. Since $N_{i,t} \ge 1$ increases by one at each test of i, there are at most $24\log T/(c_i-\theta_i)^2$ such tests. Hence,

$$\sum_{i\in[L]\setminus[L^*]}(c_i-\theta_i)\,\mathbb{E}\left[\sum_{t=2}^{T}\mathbf{1}\{\mathcal{T}_{i,t}\}\mathbf{1}\{\mathcal{E}_t^c\}\right] \le \sum_{i\in[L]\setminus[L^*]}\frac{24\log T}{c_i-\theta_i}. \tag{60}$$

Misordering optimal options. Fix $i \in [L^*]$ with $s_i \geq 1$, and let p_i be as in Lemma 8, which applies under the additional condition $\theta_k < 1$ of Theorem 1. For $j \in [s_i]$, we have $\rho_j > \rho_i$ and all costs are positive, so $n_{i,t}(\rho_j) = A_{i,j} \log t$ with $A_{i,j} := 6(1+\rho_j)^2/((\rho_j - \rho_i)^2 c_i^2)$. Define

$$\zeta_{i,j} := \max\left\{ n : A_{i,j}(\log n + \log 2) \geq \frac{p_i}{2} n \right\}, \tag{61}$$

which is the constant ζ of Lemma 8 with $A = A_{i,j}$. It is finite because $p_i > 0$. It is also at least one, so that $\log \zeta_{i,j} \geq 0$ below, because $\rho_j - \rho_i \leq \rho_j$ and $c_i \leq 1$ give $A_{i,j} \geq 6 > p_i/(2 \log 2)$. Dropping the indicator of the testing event from $K_{i,j}$ and applying Lemma 8 with $A = A_{i,j}$, we obtain, for every $j \in [s_i]$,

$$K_{i,j} \leq \mathbb{E}\left[\sum_{t=2}^{T} \mathbf{1}\{\mathcal{E}_t^c\} \mathbf{1}\{N_{i,t} \leq n_{i,t}(\rho_j)\} \right] \leq \zeta_{i,j} + \frac{2}{p_i^2} + 4L. \tag{62}$$

By (61), $\zeta_{i,j}$ satisfies $A_{i,j}(\log \zeta_{i,j} + \log 2) \geq p_i \zeta_{i,j}/2$. To apply Lemma 4 below with a logarithmic factor common to all $j \in [s_i]$, we replace $\zeta_{i,j}$ inside the logarithm by ζ_{i,s_i}. This is possible because $A_{i,j}$ is nonincreasing in ρ_j by Lemma 2 and $\rho_j \geq \rho_{s_i}$, so $A_{i,j} \leq A_{i,s_i}$ and hence $\zeta_{i,j} \leq \zeta_{i,s_i}$. Together with $A_{i,j} \log t = n_{i,t}(\rho_j) \leq 24 \log t / g_i(\rho_j)^2$ from (23), this gives

$$\zeta_{i,j} \leq \frac{2A_{i,j}(\log \zeta_{i,s_i} + \log 2)}{p_i} \leq \frac{48(\log \zeta_{i,s_i} + \log 2)}{g_i(\rho_j)^2 p_i}.$$

Substituting this into (62) and (58), and using Lemma 4 together with $\sum_{j=1}^{s_i - 1} (g_i(\rho_j) - g_i(\rho_{j+1})) + g_i(\rho_{s_i}) = g_i(\rho_1) \leq c_i$, we obtain

$$M_i \leq \frac{96(\log \zeta_{i,s_i} + \log 2)}{g_i(\rho_{s_i})\, p_i} + c_i \left(\frac{2}{p_i^2} + 4L \right). \tag{63}$$

This bound does not depend on T.

Combining (63), (59), and (60) in (57), we obtain

$$R_T \leq \sum_{\substack{i \in [L^*] \\ s_i \geq 1}} \left(\frac{96(\log \zeta_{i,s_i} + \log 2)}{g_i(\rho_{s_i})\, p_i} + c_i \left(\frac{2}{p_i^2} + 4L \right) \right) + \sum_{i \in [L] \setminus [L^*]} \frac{72 \log T}{c_i - \theta_i} + (L+1)(4L+1). \tag{64}$$

Case 2: Problem-independent bound For the problem-independent bound, we truncate each threshold at the horizon. Since $K_{i,j} \leq T$ by definition, the bound $K_{i,j} \leq n_{i,T}(\rho_j)$ and (23) give

$$K_{i,j} \leq \min\{n_{i,T}(\rho_j), T\} \leq \min\left\{ \frac{24 \log T}{g_i(\rho_j)^2}, T \right\} \qquad (j \in [s_i]).$$

Hence Lemma 5 with $x_j = g_i(\rho_j)$, $a = 24 \log T$, and $h = T$ gives $M_i \leq 2\sqrt{24T \log T}$ for every i. By Lemma 3, the expected number of tests of an inefficient option i under $\mathcal{E}_t^c$ is at most $\min\{24 \log T/(c_i - \theta_i)^2, T\}$, and $\min\{a/x^2, h\}\, x \leq \sqrt{ah}$ for $x > 0$. Therefore (57) gives

$$R_T \leq 2L\sqrt{24T \log T} + (L - L^*)\sqrt{24T \log T} + (L+1)(4L+1). \tag{65}$$

□

D Proof of Theorem 3 and the Problem-Dependent Bound of CC-UCBv2

Proof. We ignore options with $c_i = \theta_i = 0$ (Section 5.1) and index the remaining options in descending order of the extended-real efficiency. Let $L' := |\{i \in [L] : c_i = 0, \theta_i > 0\}|$, so that the first L' options are those with efficiency $+\infty$.

We first check that Lemma 6 remains valid for CC-UCBv2, where the thresholds $n_{i,t}(\rho_j)$ of Lemma 2 now include the case $\rho_j = +\infty$. On $\mathcal{E}_t^c$, we have $U_{i,t} \geq \theta_i$ and $L_{i,t} \leq c_i$, so I_t again contains all optimal options. The derivation of (48) uses only the sign of each swap term and the identity for passed options strictly more efficient than $I_t(i)$, and both remain valid for extended-real efficiencies.

If $c_{I_t(i)} > 0$ and $I_t(i)$ passes a zero-cost option, the corresponding factor $1 - \rho_{I_t(i)}/\rho_k$ equals one. If $c_{I_t(i)} = 0$, then $\rho_{I_t(i)} = +\infty$, no option is strictly more efficient than $I_t(i)$, and the maximum in (48) is zero. Consequently, the misordering terms of (42) with $i \le L'$ vanish, and only the following part remains:

$$\sum_{i=L'+1}^{L} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}] \max_{\substack{j\in[L^*]:\ \rho_j>\rho_i,\\ N_{i,t}\le n_{i,t}(\rho_j)}} g_i(\rho_j).$$

For $i > L'$, we have $c_i > 0$ and $\rho_i < \infty$, so g_i, $n_{i,t}$, and the ratios ρ_i/ρ_j in this sum are well defined. Lemma 6 bounds each remaining misordering term by Lemma 2, which also covers the CC-UCBv2 ordering rule. In particular, if a zero-cost option is placed after a positive-cost option i, then $L_{i,t} = 0$, and on $\mathcal{E}_t^c$ this happens only while $N_{i,t} \le 6\log t/c_i^2 = n_{i,t}(+\infty)$. Hence Lemma 6, and therefore (57), hold for CC-UCBv2 with the misordering terms restricted to $i > L'$.

We now follow Case 2 of Appendix C. Since (23) also covers the case $\rho_j = +\infty$, Lemma 5 bounds M_i for each $i > L'$ by $2\sqrt{24T\log T}$, and Lemma 3 bounds the inefficient-option term by $(L - L^*)\sqrt{24T\log T}$ as in Case 2. Therefore,

$$R_T \le 2(L - L')\sqrt{24T\log T} + (L - L^*)\sqrt{24T\log T} + (L+1)(4L+1). \tag{66}$$

□

Problem-dependent bound. Assume in addition that $\theta_i < 1$ for all $i \in [L]$. We show that CC-UCBv2 then attains the bound of Theorem 1,

$$R_T \le \sum_{i\in[L]:\rho_i<1} \frac{72\log T}{c_i - \theta_i} + O(1). \tag{67}$$

We follow Case 1 of Appendix C, starting from (57) with the misordering terms restricted to $i > L'$, and check its three contributions. Every inefficient option has $c_i > 0$, so its thresholds are well defined.

Misordering inefficient options. The bound (59) uses only $K_{i,j} \le n_{i,T}(\rho_j)$, (23), and Lemma 4. All three remain valid when $\rho_j = +\infty$, in which case $g_i(\rho_j) = c_i$.

Selecting inefficient options. The proof of Lemma 3 uses only that an included option satisfies $U_{i,t} \ge L_{i,t}$. Both inclusion conditions of CC-UCBv2 imply this inequality, so (60) holds.

Misordering optimal options. The proof of Lemma 8 uses the assumption of Section 4 only to show that $i \in I_t$ on $\mathcal{E}_t^c$. This also holds under CC-UCBv2. Let $i \in [L^*]$ with $i > L'$. On $\mathcal{E}_t^c$, we have $U_{i,t} \ge \theta_i$ and $L_{i,t} \le c_i$, and $\theta_i \ge c_i > 0$. Hence either $L_{i,t} = 0 < U_{i,t}$ or $U_{i,t}/L_{i,t} \ge \rho_i \ge 1$, and in both cases $i \in I_t$. The rest of that proof does not depend on the ordering rule, so Lemma 8 holds for CC-UCBv2. In the derivation of (63), a more efficient option $j \le L'$ has $n_{i,t}(\rho_j) = 6\log t/c_i^2$, so we set $A_{i,j} := 6/c_i^2$. This value still satisfies $A_{i,j} \ge 6$, and $A_{i,j} \le A_{i,s_i}$ holds because $n_{i,t}$ is nonincreasing on $(\rho_i, +\infty]$ by Lemma 2. Hence (63) holds for every $i \in [L^*]$ with $i > L'$ and $s_i \ge 1$. Options with $i \le L'$ have no misordering term.

Combining the three contributions as in Case 1 gives (67), where the $O(1)$ term is the first sum in the final bound of Case 1 restricted to $i > L'$, plus $(L+1)(4L+1)$.

E Proof of Theorem 4

Throughout this section we fix an instance and analyze BC-CC-UCBv2 (Algorithm 3). A quantity is called a constant if it may depend on the instance but not on T, and $O(1)$ denotes such a constant. Section E.1 classifies the options and identifies the comparator, and Section E.2 states the decomposition of the regret and indicates where each term is bounded; the remaining subsections carry out these bounds. We used ChatGPT to suggest the proof strategy for Theorem 4. The authors verified all arguments in this appendix.

E.1 Setup

Conventions. We use $\mathcal{F}_t$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ from Section 3. All statistics computed by the algorithm at round t, namely $N_{i,t}$, $F_{i,t}$, $\hat{\theta}_{i,t}$, $\hat{c}_{i,t}$, $U_{i,t}$, $L_{i,t}$, $V_{i,t}$, $\mathcal{Z}_t$, $\mathcal{A}_t$, b_t, $V_{*,t}$, P_t, and I_t, are

$\mathcal{F}_t$-measurable. As in Appendix B, $(X_i^{(1)}, Y_i^{(1)})$ is the initial observation of option i, its n-th test yields $(X_i^{(n+1)}, Y_i^{(n+1)})$, and these pairs are i.i.d. in n. An option moves permanently from $\mathcal{Z}_t$ to $\mathcal{A}_t$ after its first observed failure.

If two lists share a prefix P, their expected rewards differ only after P is passed: for lists P, A, and B,

$$\mu(P\|A) - \mu(P\|B) = \Big(\prod_{r\in P}(1-\theta_r)\Big)\big(\mu(A) - \mu(B)\big), \tag{68}$$

where the product is the probability that every option of P fails. Similarly, the k-th option of I_t is tested with probability $\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{I_t(k),t}\}] = \prod_{k'<k}(1-\theta_{I_t(k')})$.

Auxiliary and null options. The auxiliary option 0 has $(\theta_0, c_0) := (1, 1)$ and $\rho_0 := 1$. If $b_t \neq 0$, the algorithm plays $I_t = P_t^0\|P_t^+\|(b_t)$. If $b_t = 0$, it plays $P_t^0\|P_t^+$. Appending option 0 to this list does not change the expected reward, because reaching option 0 yields $\theta_0 - c_0 = 0$. Hence we write $I_t = P_t^0\|P_t^+\|(b_t)$ in both cases. We do not regard reaching option 0 as a test, and $N_{i,t}$ is defined only for $i \in [L]$.

An option with $\theta_i = c_i = 0$ satisfies $X_{i,t} = Y_{i,t} = 0$ almost surely, and we call it a *null* option. From now on, we delete null options from every list and exclude them from every set defined below. This loses nothing: deleting a null option from a list changes neither the reward nor the probability that any other option is tested. Moreover, the initial observation of a null option is a failure, so a null option belongs to $\mathcal{A}_t$ for every $t \geq 1$. Hence it never belongs to $\mathcal{Z}_t$ and is never chosen as the terminal candidate b_t. For non-null options we use the extended-real efficiency of Section 5.1.

The boundary and the classification of options. Let $\mathcal{S} := \{i \in [L] : \theta_i = 1\}$ be the set of sure-success options. As in Section 5.2, $c^* = \min(\{1\} \cup \{c_i : i \in \mathcal{S}\})$. Let $\mathcal{S}^* := \{i \in \mathcal{S} : c_i = c^*\}$ be the set of the cheapest sure-success options and $\mathcal{S}^- := \{i \in \mathcal{S} : c_i > c^*\}$ the set of the others. If $\mathcal{S} \neq \varnothing$, then $c^* = \min_{i\in\mathcal{S}} c_i$ because $c_i \leq 1$, so $\mathcal{S}^* \neq \varnothing$, and we fix $b^* \in \mathcal{S}^*$. If $\mathcal{S} = \varnothing$, then $c^* = 1$, and we set $b^* := 0$. Whether or not $\mathcal{S}$ is empty, $\theta_{b^*} = 1$ and $c_{b^*} = c^*$. Let $\beta := 1/c^* \geq 1$ if $c^* > 0$ and $\beta := +\infty$ if $c^* = 0$. Then β is the efficiency of b^*. With $\Delta_i^* = c_i - c^*\theta_i$, the proof uses the following sets of non-null options.

Set	Definition	Role
$\mathcal{P}^*$	$\theta_i < 1, \Delta_i^* \leq 0$	optimal prefix (before b^* in I^*)
$\mathcal{W}$	$\theta_i < 1, \Delta_i^* > 0$	below the boundary
$\mathcal{S}^*$	$\theta_i = 1, c_i = c^*$	sure-success, at the boundary
$\mathcal{S}^-$	$\theta_i = 1, c_i > c^*$	sure-success, below the boundary
$\mathcal{C}$	$\theta_i < 1, c_i < c^*$	cheap; may keep $V_{*,t}$ below c^*

Every non-null option belongs to exactly one of $\mathcal{P}^*$, $\mathcal{W}$, $\mathcal{S}^*$, and $\mathcal{S}^-$, because an option of $\mathcal{S}$ has $c_i \geq c^*$. The options with positive boundary gap are exactly those of $\mathcal{W} \cup \mathcal{S}^-$, and only these appear in the regret bound. The set $\mathcal{C} \subseteq \mathcal{P}^* \cup \mathcal{W}$ is not part of this partition. Until an option of $\mathcal{C}$ fails for the first time, it stays in $\mathcal{Z}_t$. Since it costs less than c^*, the algorithm may choose it as the terminal candidate, and $V_{*,t}$ may be smaller than c^*.

The sets $\mathcal{P}^*$ and $\mathcal{W}$ are separated by the efficiency β of the boundary: for a non-null option i with $\theta_i < 1$,

$$\Delta_i^* = c_i\Big(1 - \frac{\rho_i}{\beta}\Big) \text{ if } c_i > 0, \qquad \text{and} \qquad \Delta_i^* \leq 0 \iff \rho_i \geq \beta. \tag{69}$$

Indeed, if $c_i > 0$, then $c^* = 1/\beta$ gives $c_i - c^*\theta_i = c_i(1 - \rho_i/\beta)$, which is nonpositive if and only if $\rho_i \geq \beta$; if $c_i = 0$, then $\theta_i > 0$, $\rho_i = +\infty \geq \beta$, and $\Delta_i^* = -c^*\theta_i \leq 0$. Hence every option of $\mathcal{P}^*$ has efficiency at least β, and every option of $\mathcal{W}$ has positive cost and efficiency smaller than β. If $c^* = 0$, then $\Delta_i^* = c_i$, so every option of $\mathcal{P}^*$ has zero cost.

Reachability. Let

$$q := \prod_{i:\theta_i<1}(1-\theta_i) > 0. \tag{70}$$

Lemma 9 (Reachability). *For every $t \geq 1$, the following hold.*

(a) Every option with $\theta_i = 1$ belongs to $\mathcal{Z}_t$. Hence every option of $\mathcal{A}_t$, and in particular of P_t, satisfies $\theta_i < 1$.

(b) For every $i \in P_t$, and for $i = b_t$ when $b_t \neq 0$,

$$\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}] \geq q. \tag{71}$$

(c) If $c^* < 1$ and $t \geq 2$, then $b_t \neq 0$.

Proof. (a) An option with $\theta_i = 1$ never fails, so $F_{i,t} = 0$. (b) The options placed before an option of P_t, or before b_t, are distinct options r of P_t, which satisfy $\theta_r < 1$ by (a). The probability that i is tested is the product of their $1 - \theta_r$, which is at least q. (c) If $c^* < 1$, then $\mathcal{S}^* \neq \varnothing$. Every $b \in \mathcal{S}^*$ belongs to $\mathcal{Z}_t$ by (a) and satisfies $L_{b,t} = \max\{\hat{c}_{b,t} - u_{b,t}, 0\} < 1 = L_{0,t}$, because $\hat{c}_{b,t} \leq 1$ and $u_{b,t} > 0$ for $t \geq 2$. Hence $b_t \neq 0$. □

The comparator. The regret R_T depends on the optimal ordered subset of Section 3 only through its expected reward, so any list that maximizes μ can serve as I^*. In this section, we redefine I^* as the list given in the next lemma, which shows that this list maximizes μ.

Lemma 10 (Comparator). *Let $I^* := \mathrm{ord}(\mathcal{P}^*)\|(b^*)$, where $\mathrm{ord}(\cdot)$ is as in Appendix C, with the extended efficiency, so that zero-cost options come first. Then $\mu(I^*) \geq \mu(I)$ for every list I.*

Proof. First let $c^* = 0$. Every option of $\mathcal{P}^*$ then has zero cost, and b^* has $\theta_{b^*} = 1$ and $c_{b^*} = 0$. Hence I^* incurs no cost and succeeds with probability one, so $\mu(I^*) = 1$. Every list I satisfies $\mu(I) \leq 1$, because the reward of a round is at most one.

Now let $c^* > 0$. We use two facts.

(i) Options placed after an option with success probability one are never tested. Hence adding or deleting such options does not change μ.

(ii) Let I contain i immediately followed by j, and let I' be I with i and j exchanged. If $\rho_j \geq \rho_i$, then $\mu(I') \geq \mu(I)$, and if $\rho_j = \rho_i$, then $\mu(I') = \mu(I)$. Indeed, by (5), $\mu(I') - \mu(I)$ is a nonnegative factor times $\theta_j c_i - \theta_i c_j$. This equals $c_i c_j(\rho_j - \rho_i)$ if $c_i, c_j > 0$, and equals $\theta_j c_i \geq 0$ if $c_j = 0$, with equality if also $c_i = 0$. The case $c_i = 0 < c_j$ does not occur, because then $\rho_i = +\infty > \rho_j$.

Fix a list I. We transform I into I^* in three steps, none of which decreases μ.

Step 1: end the list with b^*. If I does not contain b^*, append b^* to the end of I. If I contains an option of $\mathcal{S}$, the end of I is never reached, so μ does not change. Otherwise, the end of I is reached with some probability p, and μ increases by $p(\theta_{b^*} - c_{b^*}) = p(1 - c^*) \geq 0$.

Step 2: add the missing options of $\mathcal{P}^*$. Append, after b^*, the options of $\mathcal{P}^*$ that are not yet in the list. By (i), μ does not change, because $\theta_{b^*} = 1$. After this step, the list contains every option of $\mathcal{P}^*$.

Step 3: sort. Sort the list by exchanges of adjacent options into descending order of efficiency. Among the options with efficiency β, place the options of $\mathcal{P}^*$ first, then b^*, and then the other options of $\mathcal{S}^*$. Each exchange moves an option of larger or equal efficiency forward, so by (ii) no exchange decreases μ.

It remains to identify the sorted list. Every option of $\mathcal{S}$ has efficiency $1/c_i \leq \beta$. By (69), an option with $\theta_i < 1$ has efficiency at least β if and only if it belongs to $\mathcal{P}^*$. Hence the options with efficiency larger than β all belong to $\mathcal{P}^*$, and the options placed before b^* are exactly the options of $\mathcal{P}^*$, in descending order of efficiency. By (i), deleting the options after b^* does not change μ. The result is $\mathrm{ord}(\mathcal{P}^*)\|(b^*)$, except possibly for the order among options of equal efficiency, which does not affect μ by (ii). Hence $\mu(I^*) \geq \mu(I)$. □

Consequently, $\mathrm{reg}_t = \mu(I^*) - \mu(I_t) \geq 0$ with this I^*, and the bound $R_T \leq L + 1 + \sum_{t=2}^{T} \mathbb{E}[\mathrm{reg}_t]$ of (21) applies.

E.2 Outline of the Proof

Compared with Appendix C, the new difficulty is the confidence bound $V_{*,t}$ on the boundary cost c^*. It can be too low while an option of $\mathcal{C}$ has not yet failed, and too high while no option of $\mathcal{S}^*$ has enough observations. We control the first difficulty by a random time τ: after τ, $V_{*,t} \geq c^*$ on $\mathcal{E}_t^c$. We control the second by an event $\mathcal{N}_t$: on $\mathcal{E}_t^c$, outside $\mathcal{N}_t$, and for $t \geq t_0$, $V_{*,t}$ is close to c^*.

Discovery. The *discovery time* τ is the first round $t \geq 2$ at which every option of $\mathcal{C}$ has failed at least once:

$$\tau := \inf\{t \geq 2 : \mathcal{C} \subseteq \mathcal{A}_t\}. \tag{72}$$

The event $\{\tau \leq t\}$ is determined by the history before round t, so τ is a stopping time. Since $\mathcal{A}_t$ is nondecreasing, $\mathcal{C} \cap \mathcal{Z}_t = \varnothing$ for every $t \geq \tau$. We say that discovery is *complete* at round t if $t \geq \tau$. In that case every option of $\mathcal{Z}_t$ has cost at least c^*, so $V_{*,t} \geq c^*$ on $\mathcal{E}_t^c$ (Lemma 13(a)). The time τ is used only in the analysis; the algorithm need not know $\mathcal{C}$ or τ.

Certification of the boundary. We say that the boundary is *certified* at round t if $V_{*,t} \leq c^* + \Delta_i^*/4$ for every $i \in \mathcal{W}$. If the boundary is certified and $\mathcal{E}_t^c$ holds, an option $i \in \mathcal{W}$ can enter P_t only while $N_{i,t}$ is small (Lemma 13(b)).

If $c^* = 1$, the boundary is always certified, because $V_{*,t} \leq V_{0,t} = 1 = c^*$. In this case we set $\mathcal{N}_t := \varnothing$ and $t_0 := 2$.

If $c^* < 1$, then $\mathcal{S}^* \neq \varnothing$, and certification requires enough observations of some option of $\mathcal{S}^*$. We define

$$\gamma := \frac{q}{8|\mathcal{S}^*|}, \qquad \mathcal{N}_t := \Big\{\max_{b \in \mathcal{S}^*} N_{b,t} < \gamma t\Big\}. \tag{73}$$

The event $\mathcal{N}_t$ means that every option of $\mathcal{S}^*$ has fewer than γt observations. On $\mathcal{N}_t^c$, some option of $\mathcal{S}^*$ has confidence radius at most $\sqrt{1.5 \log t/(\gamma t)}$. Let $t_0 \geq 2$ be a constant such that

$$2\sqrt{\frac{1.5 \log t}{\gamma t}} \leq \frac{\Delta_i^*}{4} \qquad \text{for every } i \in \mathcal{W} \text{ and every } t \geq t_0, \tag{74}$$

which exists because the left-hand side tends to zero; if $\mathcal{W} = \varnothing$, let $t_0 := 2$.

Rounds with the guarantees. Define the $\mathcal{F}_t$-measurable events

$$\mathcal{G}_t^{\mathrm{disc}} := \mathcal{E}_t^c \cap \{t \geq \tau\}, \qquad \mathcal{G}_t^{\mathrm{cert}} := \mathcal{G}_t^{\mathrm{disc}} \cap \mathcal{N}_t^c \cap \{t \geq t_0\}. \tag{75}$$

On $\mathcal{G}_t^{\mathrm{disc}}$, the confidence intervals are valid and discovery is complete. On $\mathcal{G}_t^{\mathrm{cert}}$, the boundary is also certified by Lemma 13(b).

Decomposition. By (21),

$$R_T \leq L + 1 + \mathbb{E}\Big[\sum_{t=2}^{T} \mathrm{reg}_t \mathbf{1}\{(\mathcal{G}_t^{\mathrm{cert}})^c\}\Big] + \mathbb{E}\Big[\sum_{t=2}^{T} \mathrm{reg}_t \mathbf{1}\{\mathcal{G}_t^{\mathrm{cert}}\}\Big]. \tag{76}$$

We call a round *exceptional* if $\mathcal{G}_t^{\mathrm{cert}}$ fails. The first expectation, over the exceptional rounds, is $O(1)$ by Lemma 20. For the second, Lemma 15 shows that, on $\mathcal{G}_t^{\mathrm{disc}}$ and hence on $\mathcal{G}_t^{\mathrm{cert}}$, reg_t is at most the sum of three contributions defined in Section E.3:

$$\mathrm{reg}_t \leq \underbrace{\mathrm{mis}_t}_{\text{misordering in } \mathcal{P}^*} + \underbrace{\sum_{i \in \mathcal{W} \cap P_t} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\, \mathrm{inc}_{i,t}}_{\text{inclusion of } \mathcal{W}} + \underbrace{\mathrm{term}_t}_{\text{terminal candidate}}. \tag{77}$$

The following table lists the bounds on the exceptional rounds and on these three contributions.

Contribution	Source of the loss	Total expected loss	Lemma
exceptional rounds	$\mathcal{E}_t$, $t < \tau$, $\mathcal{N}_t$, or $t < t_0$	$O(1)$	20
mis_t	misordering within $\mathcal{P}^*$	$O(1)$	21
$\mathrm{inc}_{i,t}$, $i \in \mathcal{W}$	inclusion of i in the prefix	$192 \log T/\Delta_i^*$	22
terminal, $\theta_{b_t} < 1$	b_t that is not a sure-success option	$O(1)$	23
terminal, $b_t = i \in \mathcal{S}^-$	expensive sure-success b_t	$6 \log T/\Delta_i^*$	23

E.3 Deterministic Bounds on Rounds with the Guarantees

On the good event $\mathcal{E}_t^c$ of Section 4, for every $i \in [L]$,

$$\theta_i \le U_{i,t} \le \theta_i + 2u_{i,t}, \qquad c_i - 2u_{i,t} \le L_{i,t} \le c_i, \qquad c_i \le V_{i,t} \le c_i + 2u_{i,t}, \tag{78}$$

where the truncations at 0 and 1 preserve the inequalities because $\theta_i, c_i \in [0,1]$. All results of this subsection are deterministic consequences of (78) and of the definitions of τ and $\mathcal{N}_t$.

Thresholds. Recall that $u_{i,t}^2 = 1.5 \log t / N_{i,t}$. Hence $u_{i,t} \ge x$ implies $N_{i,t} \le 1.5 \log t / x^2$; we use this throughout this subsection. The following two lemmas bound how many observations an option can have while the algorithm still makes an incorrect comparison. The first concerns the order of the prefix.

Lemma 11 (Prefix misordering threshold)**.** *Suppose that $\mathcal{E}_t^c$ holds. If $i, j \in P_t$ satisfy $c_i > 0$ and $\rho_j > \rho_i$, and i is placed before j in I_t, then $N_{i,t} \le n_{i,t}(\rho_j)$, where $n_{i,t}$ is as in Lemma 2.*

Proof. The proof of Lemma 2 uses only $\mathcal{E}_t^c$ and the rule by which CC-UCBv2 orders its list. The prefix $P_t^0 \| P_t^+$ is ordered by the same rule, so the same proof applies. □

Terminal candidates are new in BC-CC-UCBv2. The next lemma bounds the observations of the terminal candidate when a cheaper option remains in $\mathcal{Z}_t \cup \{0\}$.

Lemma 12 (Terminal threshold)**.** *Suppose that $\mathcal{E}_t^c$ holds. If $b_t = i \ne 0$ and some $j \in \mathcal{Z}_t \cup \{0\}$ satisfies $c_j < c_i$, then $N_{i,t} \le 6 \log t / (c_i - c_j)^2$.*

Proof. Since b_t minimizes $L_{\cdot,t}$ over $\mathcal{Z}_t \cup \{0\}$, (78) gives $c_i - 2u_{i,t} \le L_{i,t} \le L_{j,t} \le c_j$; for $j = 0$, the last inequality is $L_{0,t} = 1 = c_0$. Hence $u_{i,t} \ge (c_i - c_j)/2$, so $N_{i,t} \le 6 \log t / (c_i - c_j)^2$. □

The confidence bound on the boundary cost. The next lemma bounds $V_{*,t}$ and derives which options enter the prefix.

Lemma 13 (Confidence bound on the boundary cost)**.**

(a) On $\mathcal{G}_t^{\mathrm{disc}}$, $V_{*,t} \ge c^*$. If moreover $c^* > 0$, then $\mathcal{P}^* \subseteq P_t$.

(b) On $\mathcal{G}_t^{\mathrm{cert}}$, every $i \in \mathcal{W}$ satisfies $V_{*,t} \le c^* + \Delta_i^*/4$, and every $i \in \mathcal{W} \cap P_t$ satisfies $N_{i,t} \le 96 \log t / (\Delta_i^*)^2$.

Proof. (a) Recall that $V_{*,t}$ is the minimum of $V_{i,t}$ over $i \in \mathcal{Z}_t \cup \{0\}$, and that $V_{0,t} = 1 \ge c^*$. Let $i \in \mathcal{Z}_t$. If $\theta_i = 1$, then $c_i \ge c^*$. If $\theta_i < 1$, then $i \notin \mathcal{C}$ because $t \ge \tau$, so again $c_i \ge c^*$. Hence $V_{i,t} \ge c_i \ge c^*$ by (78), which proves $V_{*,t} \ge c^*$.

Let $c^* > 0$ and $i \in \mathcal{P}^*$. Then $c_i \le c^* \theta_i < c^*$, so $i \in \mathcal{C}$, and $i \in \mathcal{A}_t$ because $t \ge \tau$. Moreover, (78) and $V_{*,t} \ge c^*$ give $L_{i,t} \le c_i \le c^* \theta_i \le V_{*,t} U_{i,t}$. Hence $i \in P_t$.

(b) Let $i \in \mathcal{W}$. If $c^* = 1$, then $V_{*,t} \le V_{0,t} = 1 = c^*$. If $c^* < 1$, then, since $\mathcal{N}_t$ fails, some $b \in \mathcal{S}^*$ has $N_{b,t} \ge \gamma t$. This b belongs to $\mathcal{Z}_t$ by Lemma 9(a). Hence $V_{*,t} \le V_{b,t} \le c^* + 2u_{b,t} \le c^* + 2\sqrt{1.5 \log t / (\gamma t)} \le c^* + \Delta_i^*/4$, where the last inequality uses $t \ge t_0$.

Suppose moreover that $i \in P_t$. By the definition of P_t and (78), $c_i - 2u_{i,t} \le L_{i,t} \le U_{i,t} V_{*,t} \le (\theta_i + 2u_{i,t})(c^* + \Delta_i^*/4)$. Expanding the right-hand side and using $\theta_i, c^* \le 1$ and $\Delta_i^* \le c_i \le 1$,

$$\Delta_i^* = c_i - c^* \theta_i \le \theta_i \frac{\Delta_i^*}{4} + 2u_{i,t}\Big(1 + c^* + \frac{\Delta_i^*}{4}\Big) \le \frac{\Delta_i^*}{4} + 6u_{i,t}, \tag{79}$$

so $3\Delta_i^*/4 \le 6u_{i,t}$. Hence $u_{i,t} \ge \Delta_i^*/8$, that is, $N_{i,t} \le 96 \log t / (\Delta_i^*)^2$. □

Per-round decomposition. The decomposition compares I_t with intermediate lists that end with b^*. The next lemma bounds the loss from inserting one option in front of such a list.

Lemma 14 (Insertion)**.** *Let S be a list whose options all have positive success probability and whose last option has success probability one. Let i be an option with $c_i > 0$. Then*

$$\mu(S) - \mu((i) \| S) \le \max_{r \in S} c_i \Big(1 - \frac{\rho_i}{\rho_r}\Big). \tag{80}$$

Proof. Since the last option of S always succeeds, S succeeds almost surely. When S is played, let w_r be the probability that $r \in S$ is tested. The success probability of S is $\sum_{r\in S} w_r\theta_r = 1$, and its expected cost is $c(S) := \sum_{r\in S} w_r c_r$. Hence $\mu(S) = 1 - c(S)$, and

$$\mu(S) - \mu((i)\|S) = (1 - c(S)) - \big[\theta_i - c_i + (1-\theta_i)(1-c(S))\big] = c_i - \theta_i c(S). \tag{81}$$

Since $\theta_r > 0$, we have $c_r = \theta_r/\rho_r$ for every $r \in S$ (with $1/\rho_r = 0$ if $c_r = 0$). Hence $c(S) \geq \big(\min_{r\in S} 1/\rho_r\big)\sum_{r\in S} w_r\theta_r = \min_{r\in S} 1/\rho_r$. Since $\theta_i \geq 0$ and $\theta_i = c_i\rho_i$, we obtain $c_i - \theta_i c(S) \leq \max_{r\in S}(c_i - c_i\rho_i/\rho_r)$. □

Let $(p_1, \dots, p_n) := P_t^0\|P_t^+$ be the options of P_t in the order in which they appear in I_t. Let $\pi_t := \prod_{r\in P_t}(1-\theta_r)$ be the probability, given $\mathcal{F}_t$, that every option of P_t fails, so that the last position of I_t is reached. For $i \in P_t$, we say that j lies behind i if $j \in P_t$ is placed after i in I_t. Define

$$\mathrm{mis}_t := \sum_{i\in\mathcal{P}^*\cap P_t:\, c_i>0} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\, c_i \max_{j\in\mathcal{P}^* \text{ behind } i} \Big(1 - \frac{\rho_i}{\rho_j}\Big)_+, \tag{82}$$

$$\mathrm{inc}_{i,t} := c_i \max_{j\in\{b^*\}\cup\{j\in\mathcal{P}^*: j \text{ behind } i\}} \Big(1 - \frac{\rho_i}{\rho_j}\Big) \qquad (i \in \mathcal{W}\cap P_t), \tag{83}$$

$$\mathrm{term}_t := \pi_t\big[(1-c^*) - (\theta_{b_t} - c_{b_t})\big], \tag{84}$$

with $\max\varnothing := 0$. The *misordering loss* mis_t bounds the loss from misordering options within $\mathcal{P}^*$, the optimal prefix. For an option $i \in \mathcal{W}$, the *inclusion loss* $\mathrm{inc}_{i,t}$ bounds the loss, given that i is tested, of inserting i in front of the list formed by the options of $\mathcal{P}^*$ behind i followed by b^*. The *terminal loss* term_t is the gain from replacing b_t by b^*, multiplied by the probability π_t that the last position is reached.

Lemma 15 (Per-round decomposition)**.** *On* $\mathcal{G}_t^{\mathrm{disc}}$,

$$\mathrm{reg}_t \leq \mathrm{mis}_t + \sum_{i\in\mathcal{W}\cap P_t} \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\,\mathrm{inc}_{i,t} + \mathrm{term}_t. \tag{85}$$

Proof. By Lemma 9(a), every p_k satisfies $\theta_{p_k} < 1$, so it belongs to $\mathcal{P}^*$ or to $\mathcal{W}$. For $k \in [n+1]$, define the intermediate list

$$J^{(k)} := (p_1, \dots, p_{k-1})\|\operatorname{ord}\big(\mathcal{P}^* \cap \{p_k, \dots, p_n\}\big)\|(b^*). \tag{86}$$

The list $J^{(k)}$ keeps the first $k-1$ options of the prefix, and in the rest of the prefix it keeps only the options of $\mathcal{P}^*$, sorted by efficiency. The last list $J^{(n+1)} = (p_1, \dots, p_n)\|(b^*)$ differs from I_t only in its last option. We write $\mathrm{reg}_t = (\mu(I^*) - \mu(J^{(n+1)})) + (\mu(J^{(n+1)}) - \mu(I_t))$. Since $J^{(n+1)}$ and I_t share the prefix $(p_1, \dots, p_n)$ and $\mu((b^*)) = 1 - c^*$, (68) gives $\mu(J^{(n+1)}) - \mu(I_t) = \mathrm{term}_t$.

Next we show $\mu(J^{(1)}) = \mu(I^*)$. If $c^* > 0$, Lemma 13(a) gives $\mathcal{P}^* \subseteq P_t$ on $\mathcal{G}_t^{\mathrm{disc}}$, so $J^{(1)} = I^*$. If $c^* = 0$, every option of $J^{(1)}$ has zero cost and $J^{(1)}$ ends with b^*, so $\mu(J^{(1)}) = 1 = \mu(I^*)$. Hence $\mu(I^*) - \mu(J^{(n+1)}) = \sum_{k=1}^n (\mu(J^{(k)}) - \mu(J^{(k+1)}))$.

Fix k, let $i := p_k$, let $\mathcal{P}^*_{>k} := \mathcal{P}^* \cap \{p_{k+1}, \dots, p_n\}$ be the set of options of $\mathcal{P}^*$ behind i, and let $S := \operatorname{ord}(\mathcal{P}^*_{>k})\|(b^*)$. The lists $J^{(k)}$ and $J^{(k+1)}$ share the prefix $(p_1, \dots, p_{k-1})$. All options of this prefix fail with probability $\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]$, the probability that i is tested. We call the parts of $J^{(k)}$ and $J^{(k+1)}$ after this prefix their continuations. By (68), $\mu(J^{(k)}) - \mu(J^{(k+1)})$ is this probability times the difference between the expected rewards of the continuations.

Suppose $i \in \mathcal{W}$. Since $i \notin \mathcal{P}^*$, the list $J^{(k)}$ continues with S, and $J^{(k+1)}$ continues with $(i)\|S$. Moreover, $c_i > 0$ by (69). The options of S are b^* and the options of $\mathcal{P}^*$ behind i. They all have positive success probability: $\theta_{b^*} = 1$, and an option of $\mathcal{P}^*$ is non-null with efficiency at least $\beta > 0$. The last option of S is b^*. Hence Lemma 14 applies, and its right-hand side $\max_{r\in S} c_i(1 - \rho_i/\rho_r)$ is $\mathrm{inc}_{i,t}$ by definition. Therefore $\mu(J^{(k)}) - \mu(J^{(k+1)}) \leq \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\,\mathrm{inc}_{i,t}$.

Suppose $i \in \mathcal{P}^*$. Then $J^{(k+1)}$ continues with $(i)\|S$. Let $r_1, \dots, r_a$ be the options of $\operatorname{ord}(\mathcal{P}^*_{>k})$ with efficiency at least ρ_i; they form the beginning of this list. The list $J^{(k)}$ continues with $\operatorname{ord}(\mathcal{P}^*_{>k} \cup$

$\{i\})\|(b^*)$. Since μ does not depend on the tie-breaking inside ord, we may take this continuation to be S with i inserted right after $r_1, \dots, r_a$ (at the front if $a = 0$). Moving i backward past $r_1, \dots, r_a$ one at a time and applying (5),

$$\mu(J^{(k)}) - \mu(J^{(k+1)}) = \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}] \sum_{l=1}^{a} \Big(\prod_{l'<l} (1 - \theta_{r_{l'}}) \Big) \big(\theta_{r_l} c_i - \theta_i c_{r_l} \big). \tag{87}$$

If $c_i = 0$, every summand equals $-\theta_i c_{r_l} \le 0$, so the difference is nonpositive. If $c_i > 0$, then $\theta_{r_l} c_i - \theta_i c_{r_l} = \theta_{r_l} c_i (1 - \rho_i/\rho_{r_l})$. Moreover, $\sum_l \prod_{l'<l} (1 - \theta_{r_{l'}}) \theta_{r_l} \le 1$, because the l-th summand is the probability that r_l is the first option among $r_1, \dots, r_a$ to succeed. Hence the difference is at most $\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\, c_i \max_{r \in \mathcal{P}^*_{>k}} (1 - \rho_i/\rho_r)_+$, which is the term of i in mis_t.

Summing the bounds for $k = 1, \dots, n$, and adding term_t, proves the lemma. □

E.4 Rounds Without the Guarantees

A round is exceptional if $\mathcal{E}_t$ occurs, $t < \tau$, $\mathcal{N}_t$ occurs, or $t < t_0$. This subsection shows that the exceptional rounds contribute $O(1)$ to (76). Rounds on which $\mathcal{E}_t$ occurs are controlled by Lemma 1, which bounds the expectation of $B_\infty = \sum_{t \ge 2} \mathbf{1}\{\mathcal{E}_t\}$. For the rounds with $t < \tau$ and the rounds on which $\mathcal{N}_t$ occurs, we need two probabilistic tools.

Probabilistic tools. The first lemma bounds the number of tests from below when, on each counted round, the option of interest is tested with probability at least q. In its applications, $A_s = 1$ means that round s is counted, which must be determined by the history before round s, and $W_s = 1$ means that the option of interest is tested on round s. The counted rounds and the option may be chosen adaptively.

Lemma 16 (Adaptive Chernoff bound)**.** *Let $A_s \in \{0,1\}$ be $\mathcal{F}_s$-measurable and $W_s \in \{0,1\}$ be $\mathcal{F}_{s+1}$-measurable with $\mathbb{E}[W_s \mid \mathcal{F}_s] \ge q A_s$ for $s \ge 2$. Let $M_n := \sum_{s=2}^{n} A_s$ and $S_n := \sum_{s=2}^{n} W_s$. Then, for every $m > 0$,*

$$\mathbb{P}\Big(S_n \le \frac{q}{2} M_n,\ M_n \ge m \Big) \le e^{-qm/8}. \tag{88}$$

Proof. Let $\omega := \log 2$. Since $W_s \in \{0,1\}$, $\mathbb{E}[W_s \mid \mathcal{F}_s] \ge q A_s$, and $1 - x \le e^{-x}$, $\mathbb{E}[e^{-\omega W_s} \mid \mathcal{F}_s] = 1 - (1 - e^{-\omega}) \mathbb{E}[W_s \mid \mathcal{F}_s] \le \exp(-q(1 - e^{-\omega}) A_s)$. Since A_s is $\mathcal{F}_s$-measurable, $\Lambda_n := \exp(-\omega S_n + q(1 - e^{-\omega}) M_n)$ is a nonnegative supermartingale with $\mathbb{E}[\Lambda_n] \le 1$. On the event $\{S_n \le qM_n/2,\ M_n \ge m\}$, we have $-\omega S_n \ge -\omega q M_n/2$, so $\Lambda_n \ge \exp(qM_n(1 - e^{-\omega} - \omega/2)) \ge e^{qm/8}$, because $1 - e^{-\omega} - \omega/2 = (1 - \log 2)/2 > 1/8$. Markov's inequality completes the proof. □

Lemma 17 (Tests of options that may still be sure-success)**.** *For each option i with $\theta_i < 1$, the number $N_i^{\mathcal{Z}} := \sum_{t \ge 2} \mathbf{1}\{i \in \mathcal{Z}_t\} \mathbf{1}\{\mathcal{T}_{i,t}\}$ of tests of i while i has no observed failure satisfies $\mathbb{E}[N_i^{\mathcal{Z}}] \le 1/(1 - \theta_i)$.*

Proof. The n-th test of i observes $X_i^{(n+1)}$. While $i \in \mathcal{Z}_t$, every entry observed so far equals one, and i leaves $\mathcal{Z}_t$ permanently after its first observed failure. Hence $N_i^{\mathcal{Z}}$ is at most the number of entries of $(X_i^{(n)})_{n \ge 2}$ up to and including its first zero. This number is geometric with success probability $1 - \theta_i$, and its expectation is $1/(1 - \theta_i)$. □

Discovery. Before τ, an option of $\mathcal{C}$ may still be a terminal candidate and keep $V_{*,t}$ below c^*. Since each round has regret at most $L + 1$, it suffices to show that τ has finite expectation.

Lemma 18 (Discovery)**.** $\mathbb{E}[\tau] < \infty$.

Proof. If $\mathcal{C} = \varnothing$, then $\tau = 2$. Otherwise let $\delta := c^* - \max_{j \in \mathcal{C}} c_j > 0$. We do not yet know that τ is integrable, so we fix $n \in \mathbb{N}$ and consider $x_n := \mathbb{E}[\tau \wedge n]$, which is at most n and hence finite. We bound x_n by a constant that does not depend on n.

Every round t with $2 \le t < \tau$ has some $j \in \mathcal{C} \cap \mathcal{Z}_t$ and is of at least one of the following types.

(a) $\theta_{b_t} < 1$.

(b) $\mathcal{E}_t^c$ holds and $\theta_{b_t} = 1$. We show below that $b_t \neq 0$ on such rounds.

(c) $\mathcal{E}_t$ holds.

If $\mathcal{E}_t$ occurs, the round is of type (c); otherwise it is of type (a) or (b), according to whether $\theta_{b_t} < 1$. Rounds of types (a) and (c) may overlap, which only makes the following count larger. Hence

$$\tau \wedge n \leq 2 + \#\{\text{type-(a) rounds}\} + \sum_{i \in \mathcal{S}} m_i + \#\{\text{type-(c) rounds}\}, \tag{89}$$

where all rounds are counted in $[2, \tau \wedge n)$, and, for $i \in \mathcal{S}$, m_i is the number of type-(b) rounds with $b_t = i$. We bound the expectation of each term.

For type (a), fix i with $\theta_i < 1$. The event $\{b_t = i\}$ is $\mathcal{F}_t$-measurable and implies $i \in \mathcal{Z}_t$. On this event, i is tested with probability at least q by (71). By the tower property and Lemma 17,

$$q\,\mathbb{E}\Big[\sum_{t \geq 2} \mathbf{1}\{b_t = i\}\Big] \leq \mathbb{E}\Big[\sum_{t \geq 2} \mathbf{1}\{b_t = i\}\mathbf{1}\{\mathcal{T}_{i,t}\}\Big] \leq \mathbb{E}[N_i^{\mathcal{Z}}] \leq \frac{1}{1-\theta_i}. \tag{90}$$

Summing over i, the expected number of type-(a) rounds is at most $\sum_{i:\theta_i<1} 1/(q(1-\theta_i))$.

For type (b), we show that every $i \in \mathcal{S}$ satisfies

$$q\,\mathbb{E}[m_i] \leq \frac{6}{\delta^2} \log x_n. \tag{91}$$

The idea is as follows. On a type-(b) round, a cheap option $j \in \mathcal{C}$ is still in $\mathcal{Z}_t$, and the confidence intervals are valid. The algorithm then prefers a sure-success option i to j only while i has few observations. On each such round, i is tested with probability at least q, so these rounds stop after logarithmically many tests of i. We prove (91) in four steps.

Step 1: a type-(b) round forces few observations. Consider a type-(b) round t. Since b_t minimizes $L_{\cdot,t}$ over $\mathcal{Z}_t \cup \{0\}$ and $\mathcal{E}_t^c$ holds, $L_{b_t,t} \leq L_{j,t} \leq c_j < c^* \leq 1 = L_{0,t}$, so $b_t \neq 0$. Write $i := b_t$. Then $\theta_i = 1$, so $c_i \geq c^* \geq c_j + \delta$. Since $j \in \mathcal{Z}_t$ is cheaper than i, Lemma 12 gives

$$N_{i,t} \leq \frac{6 \log t}{(c_i - c_j)^2} \leq \frac{6 \log t}{\delta^2}. \tag{92}$$

Step 2: few tests on these rounds. Fix $i \in \mathcal{S}$, and let h_i be the number of tests of i on the rounds counted by m_i. At each such test, (92) and $t < \tau \wedge n$ give $N_{i,t} \leq 6 \log(\tau \wedge n)/\delta^2$. Since $N_{i,t} \geq 1$ increases by one at each test of i, these values of $N_{i,t}$ are distinct positive integers. Hence $h_i \leq 6 \log(\tau \wedge n)/\delta^2$.

Step 3: from rounds to tests. On a round with $b_t = i$, option i is tested with probability at least q by (71). Whether round t is counted by m_i is $\mathcal{F}_t$-measurable, because τ is a stopping time. Since at most n rounds are counted, the sums involved are finite, and the tower property gives $q\,\mathbb{E}[m_i] \leq \mathbb{E}[h_i]$.

Step 4: Jensen's inequality. Since $\log$ is concave, $\mathbb{E}[\log(\tau \wedge n)] \leq \log \mathbb{E}[\tau \wedge n] = \log x_n$. Combining Steps 2–4 gives

$$q\,\mathbb{E}[m_i] \leq \mathbb{E}[h_i] \leq \frac{6}{\delta^2}\mathbb{E}[\log(\tau \wedge n)] \leq \frac{6}{\delta^2} \log x_n, \tag{93}$$

which is (91). Summing over $i \in \mathcal{S}$, the expected number of type-(b) rounds is at most $\kappa_{\mathrm{d}} \log x_n$, where $\kappa_{\mathrm{d}} := 6|\mathcal{S}|/(q\delta^2)$.

For type (c), the number of rounds is at most B_∞.

Taking expectations in (89), we obtain

$$x_n \leq 2 + \underbrace{\sum_{i:\theta_i<1} \frac{1}{q(1-\theta_i)}}_{\text{type (a)}} + \underbrace{\kappa_{\mathrm{d}} \log x_n}_{\text{type (b)}} + \underbrace{\mathbb{E}[B_\infty]}_{\text{type (c)}}. \tag{94}$$

Let $C_{\mathrm{d}} := 2 + \sum_{i:\theta_i<1} 1/(q(1-\theta_i)) + \mathbb{E}[B_\infty]$, which does not depend on n. Then $x_n - \kappa_{\mathrm{d}} \log x_n \leq C_{\mathrm{d}}$. Since $x - \kappa_{\mathrm{d}} \log x \to \infty$ as $x \to \infty$, the set of $x \geq 1$ with $x - \kappa_{\mathrm{d}} \log x \leq C_{\mathrm{d}}$ is bounded, so (x_n) is bounded by a constant independent of n. By monotone convergence, $\mathbb{E}[\tau] = \lim_n x_n < \infty$. $\square$

Certification. It remains to show that a cheapest sure-success option is observed linearly often, so that the events $\mathcal{N}_t$ are rare.

Lemma 19 (Linear observation of the boundary)**.** $\sum_{t\geq 2}\mathbb{P}(\mathcal{N}_t) < \infty$.

Proof. If $c^* = 1$, then $\mathcal{N}_t = \varnothing$ by definition (Section E.2). Let $c^* < 1$. If $\sum_{b\in\mathcal{S}^*} N_{b,t} \geq qt/8 = |\mathcal{S}^*|\gamma t$, then some $b \in \mathcal{S}^*$ has $N_{b,t} \geq \gamma t$, so $\mathcal{N}_t$ does not occur. Hence it suffices to show that $\sum_t \mathbb{P}(\sum_{b\in\mathcal{S}^*} N_{b,t} < qt/8) < \infty$.

By Lemma 9(c), $b_s \neq 0$ for every $s \geq 2$. Apply Lemma 16 with $A_s \equiv 1$ and $W_s := \mathbf{1}\{\mathcal{T}_{b_s,s}\}$, so that $M_{t-1} = t-2$ and S_{t-1} is the number of tests of terminal candidates before round t. Each terminal candidate is tested with probability at least q by (71), so $\mathbb{P}(S_{t-1} \leq q(t-2)/2) \leq e^{-q(t-2)/8}$.

We subtract the tests spent on candidates outside $\mathcal{S}^*$ to obtain a lower bound on the observations of the cheapest sure-success options. Each test counted by S_{t-1} is of at least one of the following kinds:

- a test of an option of $\mathcal{S}^*$;
- a test of b_s with $\theta_{b_s} < 1$; since $b_s \in \mathcal{Z}_s$, there are at most $\sum_{i:\theta_i<1} N_i^{\mathcal{Z}}$ such tests in total;
- a test on an error round; there are at most B_∞ of them;
- a test of $b_s = i \in \mathcal{S}^-$ on a round where $\mathcal{E}_s^c$ holds. The cheaper option b^* never leaves $\mathcal{Z}_s$, so Lemma 12 with $j = b^*$ gives $N_{i,s} \leq 6\log s/(c_i - c^*)^2 \leq 6\log t/(c_i - c^*)^2$. Since $N_{i,s} \geq 1$ increases by one at each test of i, there are at most $\kappa_{\mathrm{b}}\log t$ such tests, where $\kappa_{\mathrm{b}} := \sum_{i\in\mathcal{S}^-} 6/(c_i - c^*)^2$.

Therefore

$$\sum_{b\in\mathcal{S}^*} N_{b,t} \geq S_{t-1} - \kappa_{\mathrm{b}}\log t - K_{\mathrm{b}}, \qquad K_{\mathrm{b}} := \sum_{i:\theta_i<1} N_i^{\mathcal{Z}} + B_\infty, \tag{95}$$

and $\mathbb{E}[K_{\mathrm{b}}] < \infty$ by Lemmas 17 and 1. All but finitely many t satisfy $t \geq 4$ and $\kappa_{\mathrm{b}}\log t \leq qt/16$. For such t,

$$\Big\{\sum_{b\in\mathcal{S}^*} N_{b,t} < \frac{qt}{8}\Big\} \subseteq \Big\{S_{t-1} \leq \frac{q(t-2)}{2}\Big\} \cup \Big\{K_{\mathrm{b}} > \frac{qt}{16}\Big\}. \tag{96}$$

Indeed, if neither event on the right occurs, then $S_{t-1} > q(t-2)/2 \geq qt/4$ because $t \geq 4$, and the previous display gives $\sum_{b\in\mathcal{S}^*} N_{b,t} > qt/4 - qt/16 - qt/16 = qt/8$. The probability of the first event on the right is at most $e^{-q(t-2)/8}$, which is summable in t. The probability of the second is also summable, because $\sum_t \mathbb{P}(K_{\mathrm{b}} > qt/16) \leq 16\mathbb{E}[K_{\mathrm{b}}]/q$. □

Exceptional rounds. Combining the three lemmas above with Lemma 1 bounds the first expectation in (76).

Lemma 20 (Exceptional rounds)**.**

$$\mathbb{E}\Big[\sum_{t=2}^{T} \mathrm{reg}_t \mathbf{1}\{(\mathcal{G}_t^{\mathrm{cert}})^c\}\Big] \leq (L+1)\Big(\mathbb{E}[B_\infty] + \mathbb{E}[\tau] + \sum_{t\geq 2}\mathbb{P}(\mathcal{N}_t) + t_0\Big) = O(1). \tag{97}$$

Proof. Since $(\mathcal{G}_t^{\mathrm{cert}})^c \subseteq \mathcal{E}_t \cup \{t < \tau\} \cup \mathcal{N}_t \cup \{t < t_0\}$ and $\mathrm{reg}_t \leq L+1$, the inequality follows from $\sum_{t\geq 2}\mathbf{1}\{t<\tau\} \leq \tau$ and $\sum_{t\geq 2}\mathbf{1}\{t<t_0\} \leq t_0$. The four terms are finite by Lemmas 1, 18, and 19 and because t_0 is a constant. □

E.5 Bounding the Three Contributions and Completing the Proof

We now bound the three contributions of Lemma 15 on $\mathcal{G}_t^{\mathrm{disc}}$ and $\mathcal{G}_t^{\mathrm{cert}}$.

Misordering within the optimal prefix. On $\mathcal{G}_t^{\mathrm{disc}}$, every positive-cost option i of $\mathcal{P}^*$ is in P_t and is tested with probability at least q, so $N_{i,t}$ grows linearly in t with high probability. Since i can be misordered only while $N_{i,t}$ is below a logarithmic threshold, the total expected misordering loss is finite.

Lemma 21 (Misordering inside the optimal prefix)**.** *For every $i \in \mathcal{P}^*$ with $c_i > 0$, $\sum_{t\geq 2} \mathbb{P}(N_{i,t} \leq qt/8) < \infty$. Consequently, $\mathbb{E}[\sum_{t\geq 2} \mathrm{mis}_t \mathbf{1}\{\mathcal{G}_t^{\mathrm{disc}}\}] = O(1)$.*

Proof. Fix $i \in \mathcal{P}^*$ with $c_i > 0$. Since $0 < c_i \leq c^*\theta_i$, we have $c^* > 0$. Apply Lemma 16 with $A_s := \mathbf{1}\{\mathcal{G}_s^{\mathrm{disc}}\}$ and $W_s := A_s \mathbf{1}\{\mathcal{T}_{i,s}\}$. On $\{A_s = 1\}$, Lemma 13(a) places i in P_s, so i is tested with probability at least q by (71). Thus $\mathbb{E}[W_s \mid \mathcal{F}_s] \geq qA_s$.

Lemma 16 counts only the rounds with $A_s = 1$, so we must control the other rounds. Since $A_s = 0$ only if $s < \tau$ or $\mathcal{E}_s$ occurs, the number of rounds $s \in [2, t-1]$ with $A_s = 0$ is at most $\tau + B_\infty$. Let $t \geq 8$, and suppose that $N_{i,t} \leq qt/8$ and $\tau + B_\infty \leq t/2$. Then $M_{t-1} \geq (t-2) - t/2 = t/2 - 2$. Moreover, S_{t-1} counts some of the tests of i before round t, so $S_{t-1} \leq N_{i,t} \leq qt/8 \leq (q/2)(t/2 - 2) \leq (q/2)M_{t-1}$, where the third inequality uses $t \geq 8$. Hence

$$\Big\{N_{i,t} \leq \frac{qt}{8}\Big\} \subseteq \Big\{S_{t-1} \leq \frac{q}{2}M_{t-1},\ M_{t-1} \geq \frac{t}{2} - 2\Big\} \cup \Big\{\tau + B_\infty > \frac{t}{2}\Big\}, \tag{98}$$

and Lemma 16 with $m = t/2 - 2$ bounds the probability of the first event, so that

$$\mathbb{P}\Big(N_{i,t} \leq \frac{qt}{8}\Big) \leq \exp\Big(-\frac{q(t/2-2)}{8}\Big) + \mathbb{P}\Big(\tau + B_\infty > \frac{t}{2}\Big). \tag{99}$$

The first term is summable in t. So is the second, because $\sum_t \mathbb{P}(\tau + B_\infty > t/2) \leq 2\,\mathbb{E}[\tau + B_\infty]$ and $\mathbb{E}[\tau + B_\infty] < \infty$ by Lemmas 18 and 1.

For the second claim, call a pair (i, j) *relevant* if $i, j \in \mathcal{P}^*$, $c_i > 0$, and $\rho_j > \rho_i$. In the term of i in mis_t, the maximum is positive only if some $j \in \mathcal{P}^*$ with $\rho_j > \rho_i$ lies behind i, and it is at most one. Bounding also $\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}] \leq 1$, we obtain

$$\mathrm{mis}_t \leq \sum_{(i,j)\text{ relevant}} c_i \mathbf{1}\{i, j \in P_t \text{ and } i \text{ is placed before } j\}. \tag{100}$$

Fix a relevant pair (i, j). As in the first claim, $c^* > 0$, so $\beta = 1/c^* \geq 1$ and $\rho_j \geq \beta \geq 1$. On $\mathcal{G}_t^{\mathrm{disc}}$, if i is placed before j, then Lemma 11 and (23) give $N_{i,t} \leq n_{i,t}(\rho_j) \leq \kappa \log t$, where κ is the maximum of $24/g_i(\rho_j)^2$ over the finitely many relevant pairs. Let t_1 be a constant such that $\kappa \log t \leq qt/8$ for all $t \geq t_1$. The rounds $t < t_1$ contribute at most $c_i t_1$ for this pair. For $t \geq t_1$, the event that $\mathcal{G}_t^{\mathrm{disc}}$ holds and i is placed before j is contained in $\{N_{i,t} \leq \kappa \log t\} \subseteq \{N_{i,t} \leq qt/8\}$. Therefore

$$\mathbb{E}\Big[\sum_{t\geq 2} \mathrm{mis}_t \mathbf{1}\{\mathcal{G}_t^{\mathrm{disc}}\}\Big] \leq \sum_{(i,j)\text{ relevant}} c_i\Big(t_1 + \sum_{t\geq 2} \mathbb{P}\Big(N_{i,t} \leq \frac{qt}{8}\Big)\Big) = O(1) \tag{101}$$

by the first claim. □

Inclusion of an option below the boundary. When $i \in \mathcal{W}$ is included, its loss $\mathrm{inc}_{i,t}$ depends on the most efficient option behind i, and a large loss forces a small $N_{i,t}$ by (103) below. We therefore bound the loss of each test of i by a function of $N_{i,t}$, instead of bounding every test by the largest possible loss.

Lemma 22 (Inclusion below the boundary)**.** *For every $i \in \mathcal{W}$,*

$$\mathbb{E}\Big[\sum_{t=2}^{T} \mathbf{1}\{\mathcal{G}_t^{\mathrm{cert}}\}\mathbf{1}\{i \in P_t\}\mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]\mathrm{inc}_{i,t}\Big] \leq \frac{192 \log T}{\Delta_i^*}. \tag{102}$$

Proof. By (69), $c_i > 0$ and $\rho_i < \beta$, so g_i and $n_{i,t}$ of Lemma 2 are defined on $[\beta, +\infty]$, and $g_i(\beta) = c_i(1 - \rho_i/\beta) = \Delta_i^*$. For each round t with $i \in P_t$, let $\rho_{i,t}^{\max}$ be the largest efficiency among b^* and the options of $\mathcal{P}^*$ behind i in I_t; it is at least β, the efficiency of b^*. Since g_i is nondecreasing, $\mathrm{inc}_{i,t} = g_i(\rho_{i,t}^{\max}) \geq g_i(\beta) = \Delta_i^* > 0$.

We claim that, on $\mathcal{G}_t^{\mathrm{cert}}$, if $i \in P_t$, then

$$N_{i,t} \le \frac{96 \log t}{\mathrm{inc}_{i,t}^2} \qquad \text{and} \qquad N_{i,t} \le \frac{96 \log t}{(\Delta_i^*)^2}. \tag{103}$$

The second inequality is Lemma 13(b). For the first, consider two cases. If $\rho_{i,t}^{\max} = \beta$, then $\mathrm{inc}_{i,t} = \Delta_i^*$, and the first inequality is the second. If $\rho_{i,t}^{\max} > \beta$, then β is finite, so $\beta \ge 1$, and some $j \in \mathcal{P}^*$ with $\rho_j = \rho_{i,t}^{\max} > \rho_i$ lies behind i. Lemma 11 and (23) give $N_{i,t} \le n_{i,t}(\rho_j) \le 24 \log t / g_i(\rho_j)^2 = 24 \log t / \mathrm{inc}_{i,t}^2$, which implies the first inequality.

Let $K_i := 96 \log T / (\Delta_i^*)^2$. By (103) and $\log t \le \log T$, on $\mathcal{G}_t^{\mathrm{cert}} \cap \{i \in P_t\}$ with $t \le T$, we have $N_{i,t} \le K_i$ and $\mathrm{inc}_{i,t} \le \sqrt{96 \log T / N_{i,t}}$. Since $\mathcal{G}_t^{\mathrm{cert}}$, $\{i \in P_t\}$, $N_{i,t}$, and $\mathrm{inc}_{i,t}$ are $\mathcal{F}_t$-measurable, the tower property bounds the left-hand side of the lemma by

$$\mathbb{E}\Big[\sum_{t=2}^{T} \mathbf{1}\{\mathcal{T}_{i,t}\} \mathbf{1}\{N_{i,t} \le K_i\} \sqrt{\frac{96 \log T}{N_{i,t}}}\Big]. \tag{104}$$

Since $N_{i,t} \ge 1$ increases by one at each test of i, the values of $N_{i,t}$ at the counted tests are distinct integers in $[1, K_i]$. Using $\sum_{n=1}^{k} n^{-1/2} \le 2\sqrt{k}$, the expectation is at most

$$\sum_{n=1}^{\lfloor K_i \rfloor} \sqrt{\frac{96 \log T}{n}} \le 2\sqrt{96 \log T \cdot K_i} = \frac{192 \log T}{\Delta_i^*}. \tag{105}$$

□

Terminal candidate. Finally, we bound the terminal loss term_t.

Lemma 23 (Terminal candidate)**.**

$$\mathbb{E}\Big[\sum_{t=2}^{T} \mathbf{1}\{\mathcal{G}_t^{\mathrm{cert}}\} \mathrm{term}_t\Big] \le \sum_{i:\theta_i<1} \frac{2}{1-\theta_i} + \sum_{i \in \mathcal{S}^-} \frac{6 \log T}{\Delta_i^*}. \tag{106}$$

Proof. Recall that $\mathrm{term}_t = \pi_t[(1-c^*) - (\theta_{b_t} - c_{b_t})]$ and $b_t \in \mathcal{Z}_t \cup \{0\}$. Since $\mathcal{Z}_t$ contains no null option, b_t is 0, an option of $\mathcal{S}^*$, an option of $\mathcal{S}^-$, or an option with $\theta_{b_t} < 1$. We treat these cases in turn.

If $b_t \in \mathcal{S}^*$, then $\theta_{b_t} = 1$ and $c_{b_t} = c^*$, so $\mathrm{term}_t = 0$. If $b_t = 0$, then $c^* = 1$ by Lemma 9(c), and $\mathrm{term}_t = 0$ because $\theta_0 = c_0 = 1$.

If $\theta_{b_t} < 1$, then $\mathrm{term}_t \le 2\pi_t$, because both $1 - c^*$ and $c_{b_t} - \theta_{b_t}$ are at most one. Fix i with $\theta_i < 1$. On $\{b_t = i\}$, option i is in the last position of I_t, so $\pi_t = \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]$, and $i \in \mathcal{Z}_t$. By the tower property and Lemma 17,

$$\sum_{t \ge 2} \mathbb{E}[\pi_t \mathbf{1}\{b_t = i\}] = \mathbb{E}\Big[\sum_{t \ge 2} \mathbf{1}\{b_t = i\} \mathbf{1}\{\mathcal{T}_{i,t}\}\Big] \le \mathbb{E}[N_i^{\mathcal{Z}}] \le \frac{1}{1-\theta_i}. \tag{107}$$

Hence the rounds with $\theta_{b_t} < 1$ contribute at most $\sum_{i:\theta_i<1} 2/(1-\theta_i)$.

If $b_t = i \in \mathcal{S}^-$, then $\mathrm{term}_t = \pi_t(c_i - c^*) = \pi_t \Delta_i^*$. In this case $\mathcal{S} \ne \varnothing$, so $b^* \in \mathcal{S}^*$, and b^* remains in $\mathcal{Z}_t$ by Lemma 9(a). Since $c_{b^*} = c^* < c_i$ and $\mathcal{G}_t^{\mathrm{cert}} \subseteq \mathcal{E}_t^c$, Lemma 12 with $j = b^*$ gives $N_{i,t} \le 6 \log t / (\Delta_i^*)^2 \le 6 \log T / (\Delta_i^*)^2$. As in the previous case, $\pi_t = \mathbb{E}_t[\mathbf{1}\{\mathcal{T}_{i,t}\}]$ on $\{b_t = i\}$, so by the tower property the contribution of i is at most Δ_i^* times the expected number of tests of i with $N_{i,t} \le 6 \log T / (\Delta_i^*)^2$. Since $N_{i,t} \ge 1$ increases by one at each test of i, there are at most $6 \log T / (\Delta_i^*)^2$ such tests, so this contribution is at most $6 \log T / \Delta_i^*$. □

Completing the proof.

Proof of Theorem 4. In (76), the exceptional rounds contribute $O(1)$ by Lemma 20. On $\mathcal{G}_t^{\mathrm{cert}} \subseteq \mathcal{G}_t^{\mathrm{disc}}$, Lemma 15 bounds reg_t by the three contributions. Since $\mathrm{mis}_t \ge 0$, its sum over the rounds

of $\mathcal{G}_t^{\mathrm{cert}}$ is at most its sum over the rounds of $\mathcal{G}_t^{\mathrm{disc}}$, which Lemma 21 bounds. Lemmas 22 and 23 bound the sums of the other two contributions over the rounds of $\mathcal{G}_t^{\mathrm{cert}}$ directly. Hence

$$R_T \le \sum_{i\in\mathcal{W}} \frac{192\log T}{\Delta_i^*} + \sum_{i\in\mathcal{S}^-} \frac{6\log T}{\Delta_i^*} + O(1) \le \sum_{i\in[L]:\Delta_i^*>0} \frac{192\log T}{\Delta_i^*} + O(1), \tag{108}$$

because $6 \le 192$ and $\mathcal{W}\cup\mathcal{S}^- = \{i\in[L] : \Delta_i^* > 0\}$ (a null option has $\Delta_i^* = 0$). This is the first statement. If $\theta_i < 1$ for all i, then $\mathcal{S} = \varnothing$, $c^* = 1$, and $\Delta_i^* = c_i - \theta_i$. Then $\Delta_i^* > 0$ if and only if $\rho_i < 1$, which gives the second statement. □

F Placing Options with Zero Observed Cost First

Under CC-UCBv2, a positive-cost option i belongs to I_t^0 while $L_{i,t} = 0$, and I_t^0 is ordered by $U_{i,t}$. Hence i can be placed before a zero-cost option until $N_{i,t}$ exceeds the threshold $n_{i,t}(+\infty) = 6\log t/c_i^2$ of Lemma 2, which is large when c_i is small. Since costs take values in $[0,1]$, a zero-cost option never incurs a cost, whereas a positive-cost option incurs a positive cost after a few tests. This appendix describes a refinement of the ordering rule that uses this observation, explains why it separates zero-cost options faster, and illustrates its effect numerically.

Refined ordering rule. Let $I_t^{00} := \{i \in I_t : \hat{c}_{i,t} = 0\}$. Since $\hat{c}_{i,t} = 0$ implies $L_{i,t} = 0$, we have $I_t^{00} \subseteq I_t^0$. The refined rule orders I_t^{00} and $I_t^0 \setminus I_t^{00}$ in descending order of $U_{i,t}$ and tests the list $I_t^{00}\|(I_t^0 \setminus I_t^{00})\|I_t^+$. The inclusion rule is that of CC-UCBv2.

Guarantees and faster separation. The refinement changes only the order within I_t^0, and the analysis of CC-UCBv2 does not use this order. Indeed, for $i \in I_t^0$, Lemma 2 bounds $N_{i,t}$ by $6\log t/c_i^2$ using only $L_{i,t} = 0$. Theorem 3 and (67) therefore hold for the refined rule. The refinement separates zero-cost options faster in expectation. Since $Y_{i,t} \in [0,1]$, each test of an option with $c_i > 0$ incurs a positive cost with probability at least c_i. Such an option thus leaves I_t^{00} after fewer than $1/c_i$ tests in expectation. From then on, it is placed behind every zero-cost option, because a zero-cost option always has $\hat{c}_{j,t} = 0$. Under CC-UCBv2, it can stay ahead of a zero-cost option until $N_{i,t}$ exceeds $6\log t/c_i^2$. Each such test loses at most c_i, so the total loss drops from order $\log T/c_i$ to less than one. The order of the regret bounds is unchanged.

Numerical illustration. We compare CC-UCBv2 with the refined rule on the instance

$$\boldsymbol{\theta} = (0.5, 0.6, 0.7, 0.8, 0.9, 0.5), \qquad \boldsymbol{c} = (0.1, 0.1, 0.1, 0.1, 0.1, 0)$$

for options A–F, with the settings of Section 7. Option F has zero cost, and A–E have efficiencies $5,\dots,9$, so every option is optimal and the optimal list is $(\mathrm{F},\mathrm{E},\mathrm{D},\mathrm{C},\mathrm{B},\mathrm{A})$. Figure 3 shows the result. Under CC-UCBv2, the options A–E remain in I_t^0 during the early rounds and can be placed before F, which causes the steep initial increase of the regret. The refined rule moves each of them behind F after its first positive cost and avoids most of this initial loss. The mean regret at $T = 200{,}000$ is 1413 (standard deviation 103) for CC-UCBv2 and 1189 (75) for the refined rule. The remaining regret of both comes mainly from ordering A–E, whose efficiencies are close to each other.

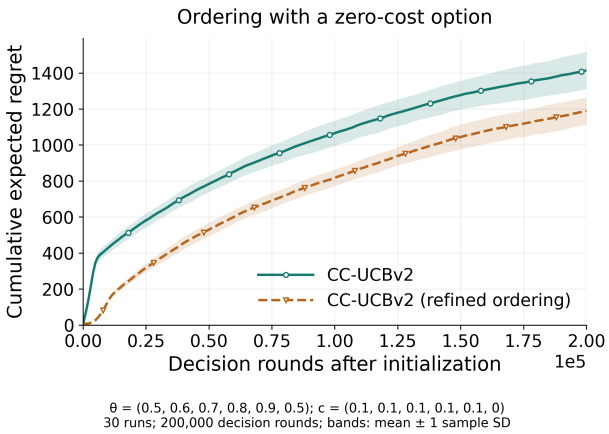

Figure 3: CC-UCBv2 and its refined ordering rule on an instance with a zero-cost option. Curves and bands use the same convention as Figure 1.

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 regret bounds for CC-UCB, the two extensions CC-UCBv2 and BC-CC-UCBv2 that remove its assumptions, the minimax lower bound, and the scope of the numerical experiments. These claims are supported by the theorems, their proofs in the main text and the appendix, and the 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 gap between the problem-independent upper and lower bounds, the fact that the match with the lower bound of Zhou et al. holds only when no sure-success option exists, the missing problem-independent bound for BC-CC-UCBv2, the larger constant in the bound for BC-CC-UCBv2, the limitations of UCB-type confidence intervals, the limited scope of the numerical experiments, and broader impacts.

   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, the algorithm descriptions, and the theorem statements. Theorems 1–4 have proof sketches in the main text and full proofs in the appendix, and the proof of Theorem 5 is given in the main text. The auxiliary lemmas are stated, proved, 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, the horizon, the number of independent runs, the exploration constant, and the compared algorithms. The supplementary code includes a reproduction script and a 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 supplementary code with independent implementations of the algorithms and a reproduction script for the reported simulation figures. The code is available at `https://github.com/tamakoshi2001/neurips2026-cost-aware-cascading`. 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, the horizon, the number of independent runs, the exploration constant, the compared algorithms, and the values of the lower-bound parameter used by CC-UCB. The supplementary code provides the exact reproduction script 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: Curves show the mean regret over 30 independent runs, and shaded bands show one sample standard deviation across runs, with the displayed lower band clipped at zero. Regret is computed from the true means, so it excludes realized-reward noise, and the algorithms share potential outcomes within each experiment. The supplementary code computes these quantities directly from the 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 CPU-based simulations reproducible on a standard laptop or desktop. The supplementary README further notes that no GPU, cluster, or specialized hardware is required. The numerical experiments section also specifies the horizon and the number of independent runs 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.

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 broader impacts of cost-aware sequential testing, including reduced testing costs in applied systems and risks from deploying the model in high-stakes domains without validating its 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 script, and license. It 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: [Yes]

    Justification: We used ChatGPT to suggest the proof strategy for Theorem 4 for BC-CC-UCBv2 in Appendix E. The authors verified all arguments in the proof. LLMs are not used in the algorithms or the experiments, and 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.