BATCHED COMBINATORIAL SEMI-BANDITS PROBLEM

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we study stochastic combinatorial semi-bandits with batched feedback, referred to as the batched combinatorial semi-bandit problem. For linear reward functions, Kim et al. (2025) showed that $O(m \log \log T)$ batches are sufficient to achieve near-optimal worst-case regret, where m is the number of base arms, and T is the time horizon. We propose `SE-BC`, which achieves near-optimal worst-case regret for general non-linear reward functions while removing the factor m from the batch complexity. We also prove a matching lower bound showing that any algorithm attaining worst-case optimal regret must use at least $\Omega(\log \log T)$ batches. Therefore, the batch complexity of `SE-BC` is optimal up to constant factors. Simulation results show that `SE-BC` achieves lower regret while using fewer batches.

1 INTRODUCTION

Sequential decision-making with combinatorial actions is a central theme in online learning. In this setting, a learner repeatedly selects a structured action, observes feedback from its selected components, and uses the accumulated observations to improve subsequent decisions. Such problems arise in a broad range of applications, including online advertising (Chen et al., 2016), online social analysis (Wen et al., 2017; Chen et al., 2016), recommendation and learning-to-rank (Kveton et al., 2015a; Li et al., 2016; Wen et al., 2015), network routing (Kveton et al., 2015b; Li et al., 2016), and wireless channel allocation (Gai et al., 2010).

Combinatorial **M**ulti-**A**rmed **B**andits (CMAB) provide a canonical model for this setting. A CMAB instance consists of m *base arms* and a family of feasible super arms $\mathcal{I} \subseteq 2^{[m]}$. At each round t, the learner chooses a *super arm* $S_t \in \mathcal{I}$, namely a non-empty subset of base arms, and then receives feedback generated by the chosen subset. In this paper, we consider the semi-bandit feedback setting, where the learner observes the individual outcome of every base arm included in the selected super arm at the end of the round. For each base arm i, its outcome is bounded in $[0, 1]$ and has mean μ_i. We write $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$ for the unknown mean vector.

Two standard approaches to CMABs are upper confidence bound (UCB) (Chen et al., 2013) and Thompson Sampling (Wang & Chen, 2018). Although these algorithms enjoy strong regret guarantees in various CMAB settings (Kveton et al., 2015c; Chen et al., 2016; Wang & Chen, 2018), they typically rely on a fully sequential interaction protocol, in which the learner updates its decision rule after every round using the latest feedback. This requirement can be impractical when interactions with the environment are limited. For example, *in online influence maximization*, a platform selects a set of seed users to initiate the diffusion of information, advertisements, or product recommendations over a social network. The selected seed set forms a combinatorial action, and the resulting cascade provides partial feedback about the activated nodes or edges (Chen et al., 2016; Wen et al., 2017). However, the diffusion data are typically collected and processed only after the campaign period ends, making fully sequential policy updates time-consuming and impractical. Similarly, in a cognitive radio network, the system assigns users to available channels, and each user-channel pair has an unknown throughput distribution. A feasible allocation, such as a matching between users and channels, forms a combinatorial action, and the throughput observed on the selected user-channel pairs provides component-level feedback (Gai et al., 2010; 2012). Since such allocation decisions are made repeatedly over many short decision periods, updating the policy after every single allocation can be computationally burdensome in large-scale systems.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This motivates our study of the **B**atched CMAB (B-CMAB). Let $0 = t_0 < t_1 < \cdots < t_L = T$ denote the batch endpoints. During batch b, the learner commits to a sequence of super arms $S_{t_{b-1}+1}, \ldots, S_{t_b} \in \mathcal{I}$ using only observations from previous batches; the same super arm may be played more than once. The feedback generated in the batch, $\{Z_i(t) : t_{b-1} < t \leq t_b, i \in S_t\}$, is revealed only after time t_b.

Despite the practical motivation of B-CMAB, the optimal batch complexity-regret tradeoff for B-CMAB remains poorly understood. The most closely related result is due to Kim et al. (2025), who showed that, for linear reward functions, $O(m \log \log T)$ batches are sufficient to achieve the near-optimal worst-case regret bound $\widetilde{O}(\sqrt{mkT})$, where k is the maximum size of a super arm. When compared with the batched multi-armed bandit and batched linear bandit settings, where $O(\log \log T)$ batches are sufficient to achieve near-optimal worst-case regret (Perchet et al., 2016; Gao et al., 2019; Ren et al., 2024), the $O(m \log \log T)$ batch complexity in the existing result reveals an additional factor m. This raises an open question:

Can we achieve near-optimal worst-case regret in B-CMAB using only $O(\log \log T)$ batches?

On the lower-bound side, $\Omega(\log \log T)$ batches are known to be necessary for achieving minimax-optimal regret in classical batched multi-armed bandits and batched linear bandits (Gao et al., 2019; Perchet et al., 2016). These lower bounds, however, do not directly extend to B-CMAB. In combinatorial semi-bandits, a single action may reveal observations from multiple base arms simultaneously, and this richer feedback could potentially reduce the amount of batches required. It is therefore unclear whether the same $\Omega(\log \log T)$ barrier persists in the combinatorial semi-bandit setting. This leads to the second open question:

Must any B-CMAB algorithm that achieves worst-case optimal regret use $\Omega(\log \log T)$ batches?

In this paper, we answer both questions affirmatively. Our main contributions are as follows:

- We propose `SE-BC`, an elimination-based algorithm that achieves near-optimal worst-case pseudo-α-regret using only $O(\log \log T)$ batches under monotonicity and additive smoothness.
- We prove a matching batch-complexity lower bound, showing that any policy with near-optimal worst-case regret must use $\Omega(\log \log T)$ batches.
- Simulations show that `SE-BC` achieves lower regret than `CUCB` and `AROQ` with fewer batches.

For our upper bound, we depart from prior methods (Kim et al., 2025) that track exploration at the level of individual base arms and introduce a new measure of exploration at the super-arm level. Specifically, we construct a dyadic deficiency score that aggregates the pull-count deficits of the base arms in a super arm and quantifies how well that super arm has been explored. This score can be updated within a batch without reward feedback and is used to guide the allocation of pulls toward super arms that remain insufficiently explored. For our lower bound, we combine the disjoint-path construction (Kveton et al., 2015c) with the crossing-event argument (Gao et al., 2019) to capture the difficulty of batched learning in the combinatorial setting. Specifically, we construct a linear-reward Bernoulli subclass with disjoint k-path super arms and analyze it over adaptive batch grids. Together, our upper and lower bounds show that `SE-BC` achieves near-optimal worst-case regret with the optimal order of batch complexity.

Table 1 compares the regret and feedback-batch guarantees of representative MAB and CMAB algorithms.

2 Related Work

Early CMAB studies considered linear rewards with semi-bandit feedback (Gai et al., 2010; 2012), followed by general frameworks covering nonlinear rewards and probabilistically triggered arms (Chen et al., 2013; 2016). Subsequent work refined the smoothness conditions to reduce the dependence of regret on the number of selected or triggered arms (Merlis & Mannor, 2019; Liu et al., 2022).

Table 1: Main regret terms and feedback-batch complexity for CMAB.

Assumptions	Algorithm / Result	Regret	Batch complexity
Linear	`AROQ-CMAB` (Kim et al., 2025)	$\widetilde{O}(\sqrt{mkT})$	$O\big(m \log\log(Tk/m)\big)$
	`SROQ-CMAB` (Kim et al., 2025)	$\widetilde{O}(k\sqrt{mT})$	$\Theta(\log\log T)$
Assumptions 1–3	`CUCB` (Wang & Chen, 2017)	$\widetilde{O}(\kappa\sqrt{mkT})$	$\Theta(T)$
	`SE-BC` (Our work)	$\widetilde{O}(\kappa\sqrt{mkT})$	$O(\log\log T)$
	Lower bound (Theorem 2)	$\sqrt{mkT}\,\mathrm{polylog}(m,T)$	$\Omega(\log\log T)$

UCB and Thompson Sampling are two standard approaches within these frameworks. UCB-based methods have established sharp regret guarantees for semi-bandits (Chen et al., 2013; 2016; Combes et al., 2015). Combinatorial Thompson Sampling has been analyzed for general reward functions (Wang & Chen, 2018), with subsequent work sharpening its regret bounds and their dependence on action size and dimension (Perrault et al., 2020; Zhang & Combes, 2024). However, these studies primarily assume fully sequential feedback, allowing observations from each round to inform subsequent decisions.

Batched bandits instead allow newly collected feedback to be used only at batch boundaries. For classical MAB, Perchet et al. (2016); Gao et al. (2019) characterized the regret–batch tradeoff, showing that $\Theta(\log\log T)$ batches are necessary and sufficient to achieve minimax-optimal regret up to logarithmic factors. Jin et al. (2021) studied the anytime setting, where the horizon is not known in advance, while Jin et al. (2020) investigated the optimal batch complexity for asymptotically optimal regret. Beyond classical MAB, Ren et al. (2024); Hanna et al. (2023) established $O(\log\log T)$-batch guarantees for linear and contextual linear bandits. Batched learning has also been extended to contextual slate bandits with generalized-linear rewards and scalar feedback (Goyal et al., 2026).

For batched combinatorial semi-bandits, Åkerblom et al. (2023) analyzed the Bayesian regret of Thompson Sampling with fixed-length batches. Kim et al. (2025) developed algorithms with infrequent updates. Viewing their update intervals as batches, `AROQ-CMAB` achieves $\widetilde{O}(\sqrt{mkT})$ regret using $O(m\log\log(Tk/m))$ batches for linear rewards, while `SROQ-CMAB` achieves $\widetilde{O}(k\sqrt{mT})$ regret using $\Theta(\log\log T)$ batches. Our work studies the regret–batch tradeoff under mild assumptions, providing an $O(\log\log T)$-batch algorithm that achieves near-optimal worst-case regret, together with a matching lower bound on the number of batches required to attain such regret.

3 Problem Setting

3.1 Stochastic CMAB with semi-bandit feedback

Let $[m] \coloneqq \{1, 2, \ldots, m\}$ be the set of base arms, where m is the total number of base arms. The family of feasible super arms is a nonempty set $\mathcal{I} \subseteq 2^{[m]}$. Define $k \coloneqq \max_{S\in\mathcal{I}} |S|$, so every feasible super arm contains at most k base arms. At each round t, each base arm $i \in [m]$ generates a random outcome $Z_i(t)$ with marginal distribution supported in $[0, 1]$ and mean μ_i. The outcomes are independent across time for each fixed arm, but we do not assume that $Z_1(t), \ldots, Z_m(t)$ are independent within the same round; arbitrary contemporaneous correlations across base arms are allowed. We additionally assume $\mathbb{E}[Z_i(t) \mid \mathcal{F}_{t-1}] = \mu_i$ for every i and t, where $\mathcal{F}_t$ is generated by the outcome vectors through round t and the learner's independent internal randomness. This condition ensures that past observations do not change the conditional mean of a future outcome; it holds, in particular, when the outcome vectors are independent across rounds.

Recall that t_b denotes the endpoint of the b-th batch and L is the total number of batches used within T rounds. For any $b \in [L]$ and $t \in (t_{b-1}, t_b]$, based on the observations collected before round t_{b-1}, the learner selects a super arm $S_t \in \mathcal{I}$. The expected reward of S_t is denoted by $r_{\boldsymbol{\mu}}(S_t)$. At the end of round t_b, the learner receives semi-bandit feedback of S_t, namely the outcomes of the selected

base arms: $\{(i, Z_i(t)) : i \in S_t\}$. To analyze regret under this feedback model, we make the following assumptions on the expected reward function.

3.2 Assumptions and Regret

We assume the following three properties, which are standard in the literature (Chen et al., 2013; 2016; Wang & Chen, 2018).

Assumption 1 (Monotonicity). For every $S \in \mathcal{I}$ and any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in [0,1]^m$ satisfying $\theta_i \leq \theta'_i$ for all $i \in [m]$, we have $r_{\boldsymbol{\theta}}(S) \leq r_{\boldsymbol{\theta}'}(S)$.

Assumption 2 (Additive smoothness). There exists a constant $\kappa \geq 1$ such that for every $S \in \mathcal{I}$ and every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in [0,1]^m$, $|r_{\boldsymbol{\theta}}(S) - r_{\boldsymbol{\theta}'}(S)| \leq \kappa \sum_{i \in S} |\theta_i - \theta'_i|$.

Assumption 3 (Offline α-approximation oracle). For a fixed $\alpha \in (0,1]$, we have access to an offline oracle Oracle_α that, for every $\boldsymbol{\theta} \in [0,1]^m$, returns a feasible super arm $\mathrm{Oracle}_\alpha(\boldsymbol{\theta}) \in \mathcal{I}$ satisfying $r_{\boldsymbol{\theta}}\big(\mathrm{Oracle}_\alpha(\boldsymbol{\theta})\big) \geq \alpha \max_{S \in \mathcal{I}} r_{\boldsymbol{\theta}}(S)$.

These assumptions are satisfied by many commonly used reward models in combinatorial bandits, including linear rewards (Gai et al., 2012; Kveton et al., 2015b) and probabilistic maximum coverage rewards (Chen et al., 2016).

Example 1 (Linear reward). Consider the linear reward model, where the reward obtained from a super arm S is the sum of the outcomes of its selected base arms: $R(S, Z) = \sum_{i \in S} Z_i$. If the mean outcome of arm i is μ_i, then $r_{\boldsymbol{\mu}}(S) = \mathbb{E}[R(S, Z)] = \sum_{i \in S} \mathbb{E}[Z_i] = \sum_{i \in S} \mu_i$. Therefore, the expected reward depends only on the mean vector $\boldsymbol{\mu}$ of the selected arms. In particular, this conclusion does not require the outcomes $\{Z_i : i \in S\}$ to be independent. The monotonicity condition is immediate, and the additive smoothness condition holds with $\kappa = 1$.

Example 2 (Probabilistic maximum coverage). Probabilistic maximum coverage is a standard nonlinear reward model, often used to describe influence maximization. In the simplest case, each selected base arm corresponds to a seed node, and $Z_i \in \{0,1\}$ indicates whether node i successfully activates a target. The reward is one if at least one selected node succeeds, and zero otherwise. Hence $R(S, Z) = 1 - \prod_{i \in S}(1 - Z_i)$. Under the standard independent activation model, where the activation probabilities are given by $\boldsymbol{\mu}$, the expected reward is $r_{\boldsymbol{\mu}}(S) = \mathbb{E}[R(S, Z)] = 1 - \prod_{i \in S}(1 - \mu_i)$. This reward is monotone in $\boldsymbol{\mu}$. Moreover, changing the mean of arm i can change the expected reward by at most $|\mu_i - \mu'_i|$, and thus the additive smoothness condition holds with $\kappa = 1$.

Define $R_{\min} \coloneqq \min_{S \in \mathcal{I}} r_{\mathbf{0}}(S), R_{\max} \coloneqq \max_{S \in \mathcal{I}} r_{\mathbf{1}}(S), \Delta_{\max} \coloneqq R_{\max} - R_{\min}$, where $\mathbf{0} \coloneqq (0, \ldots, 0)$ and $\mathbf{1} \coloneqq (1, \ldots, 1)$. By monotonicity, $R_{\min} \leq r_{\boldsymbol{\theta}}(S) \leq R_{\max}, \forall \boldsymbol{\theta} \in [0,1]^m, \forall S \in \mathcal{I}$. Let $S^\star \in \arg\max_{S \in \mathcal{I}} r_{\boldsymbol{\mu}}(S)$ be an optimal super arm and write $\mathrm{OPT} = r_{\boldsymbol{\mu}}(S^\star)$. We assume that $\mathrm{OPT} \geq 0$, while allowing individual super arms to have negative expected rewards. We define the pseudo-α-regret as

$$R_\alpha(T) \coloneqq \sum_{t=1}^{T} \big(\alpha \mathrm{OPT} - r_{\boldsymbol{\mu}}(S_t)\big).$$

4 High Level Ideas

The algorithms of Kim et al. (2025) illustrate an important limitation of controlling exploration at the level of individual base arms. In `AROQ-CMAB`, each base arm maintains its own epoch, and the UCB indices are updated when the pull count of some base arm reaches a prescribed threshold. As a result, the update schedule is organized separately across the m base arms, leading to the additional factor m in the batch complexity. `SROQ-CMAB` synchronizes these updates and reduces the number of batches to $\Theta(\log \log T)$, but it still constructs and explores representative actions for individual base arms and consequently incurs a larger regret dependence on the super-arm size. Thus, the main difficulty is not UCB itself, but rather that the progress of exploration is tracked and controlled.

Our key idea is to change the unit of exploration from a base arm to a *super arm*. Under batched feedback, reward estimates cannot be updated within a batch, but the pull counts of the base arms are always known. These counts already tell us how much information has been collected and how

much uncertainty remains. We therefore use them directly to measure whether a super arm still needs exploration.

More specifically, in batch ℓ, our goal is to collect enough information so that super arms whose pseudo-α gaps are larger than a target accuracy ϵ_ℓ can be eliminated at the end of the batch. Reaching this accuracy requires each base arm to be sampled sufficiently often. We therefore associate each base arm with a collection of target sampling levels and define a *dyadic deficiency score* for every super arm by aggregating the sampling deficits of its constituent base arms. Intuitively, the score measures how much uncertainty remains in that super arm relative to the current target accuracy: a larger score means that the super arm contains more insufficiently explored base arms, and pulling it reveals more information about the still unknown environment. Importantly, the score depends only on pull counts, so it can be updated immediately after every action even though reward feedback is delayed until the end of the batch.

Our within-batch strategy is simple: we keep exploring active super arms with large deficiency scores. Each pull increases the counts of all selected base arms and reduces the remaining exploration deficit. We show that once all active scores fall below the threshold, the confidence uncertainty is small enough at the ϵ_ℓ scale. After the batch feedback is revealed, we update the confidence bounds and eliminate clearly suboptimal super arms, so that every surviving arm has pseudo-α gap at most ϵ_ℓ. Repeating this process over a doubly logarithmic schedule yields near-optimal worst-case regret with only $O(\log \log T)$ batches.

5 Main Results

5.1 The Proposed Algorithm

Our algorithm is elimination-based. In each batch, it eliminates suboptimal super arms whose upper confidence bound are clearly below the verified threshold. As the number of batches increases, the elimination threshold becomes progressively smaller. Our algorithm consists of three main parts. The first part is the batch size. The second part is the confidence bound and the elimination rule, where we use the confidence bound to rule out suboptimal super arms. The last part is how to efficiently pull super arms so that the confidence radius converges at a fast rate. Finally, we combine these components to present the full algorithm. We now begin with the first part.

Batch Length. We use an explicit doubly logarithmic grid, which has also been used in MAB (Perchet et al., 2016; Gao et al., 2019) and linear bandits (Ren et al., 2024). Specifically, we divide the horizon into $L = \lceil \log_2 \log_2 T \rceil + 1$ batches with

$$t_0 = 0, t_\ell = \left\lfloor T^{1-2^{-\ell}} \right\rfloor \text{ for } \ell = 1, \ldots, L-1, \text{ and } t_L = T$$

Batch ℓ therefore consists of rounds $t_{\ell-1}+1, \ldots, t_\ell$ and has length $B_\ell = t_\ell - t_{\ell-1}$.

Confidence Radius, Verctor, and Box. Before we define the confidence radius, we introduce some notation. For each arm $i \in [m]$, let $N_{t,i} := \sum_{s=1}^{t} \mathbb{1}\{i \in S_s\}$ denote its cumulative pull count through round t, with $N_{0,i} = 0$. Write $\mathbf{N}_t = (N_{t,1}, \ldots, N_{t,m})$ for the count vector, so $\mathbf{N}_{t_\ell}$ records the counts at the end of batch ℓ. After receiving feedback from each nonterminal batch $\ell \in \{1, \ldots, L-1\}$, we update the reward estimates using all observations collected so far. For each arm i with $N_{t_\ell,i} \geq 1$, define the empirical mean $\widehat{\mu}_{\ell,i} := \frac{1}{N_{t_\ell,i}} \sum_{t=1}^{t_\ell} \mathbb{1}\{i \in S_t\} Z_i(t)$ and the confidence radius $\beta_{\ell,i} := \sqrt{\log(4mT/\delta)/(2N_{t_\ell,i})}$. For $N_{t_\ell,i} = 0$, set $\widehat{\mu}_{\ell,i} = 0$ and $\beta_{\ell,i} = 1$. Initialize $\underline{\mu}_{0,i} = 0$ and $\overline{\mu}_{0,i} = 1$ for all $i \in [m]$. After each nonterminal batch ℓ, define the coordinate-wise confidence envelopes

$$\underline{\mu}_{\ell,i} := \max\{\underline{\mu}_{\ell-1,i}, \widehat{\mu}_{\ell,i} - \beta_{\ell,i}\}, \quad \overline{\mu}_{\ell,i} := \min\{\overline{\mu}_{\ell-1,i}, \widehat{\mu}_{\ell,i} + \beta_{\ell,i}\}.$$

The corresponding vector envelopes are $\underline{\boldsymbol{\mu}}_\ell = (\underline{\mu}_{\ell,1}, \ldots, \underline{\mu}_{\ell,m})$ and $\overline{\boldsymbol{\mu}}_\ell = (\overline{\mu}_{\ell,1}, \ldots, \overline{\mu}_{\ell,m})$. Starting from $\mathcal{M}_0 = [0,1]^m$, the confidence box after batch ℓ is

$$\mathcal{M}_\ell := \{\boldsymbol{\nu} \in [0,1]^m : \underline{\mu}_{\ell,i} \leq \nu_i \leq \overline{\mu}_{\ell,i} \text{ for all } i \in [m]\}.$$

Elimination within the Batch. For each nonterminal batch $\ell \in \{1, \dots, L-1\}$, let $\widehat{S}_\ell \in \mathrm{Oracle}_\alpha(\underline{\boldsymbol{\mu}}_\ell)$. We maintain a running benchmark by setting $\Gamma_0 \coloneqq 0$ and

$$\Gamma_\ell \coloneqq \max\{\Gamma_{\ell-1}, r_{\underline{\boldsymbol{\mu}}_\ell}(\widehat{S}_\ell)\}.$$

Thus, Γ_ℓ records the largest reward certified by the lower confidence bounds up to batch ℓ, with 0 as the initial benchmark. Starting from $\mathcal{C}_0 = \mathcal{I}$, after each nonterminal batch ℓ, we eliminate any super arm whose upper-confidence reward falls below Γ_ℓ and define the remaining active family as

$$\mathcal{C}_\ell \coloneqq \{S \in \mathcal{I} : r_{\overline{\boldsymbol{\mu}}_\ell}(S) \geq \Gamma_\ell\}. \tag{1}$$

Although this definition gives a simple elimination rule, directly constructing $\mathcal{C}_\ell$ can be computationally expensive, since the number of feasible super arms in $\mathcal{I}$ may be exponential in m. We discuss how this optimization can be implemented efficiently with greedy procedure in Section C.

Explore with dyadic deficiency score. The previous part specifies the elimination rule, but does not yet determine which super arms are pulled within each batch. We now describe the within-batch pulling strategy. Let $\mathcal{S}_{\mathrm{dyad}} \coloneqq \{2^j : 0 \leq j \leq \lceil \log_2(4k) \rceil\}$. For each nonterminal batch $\ell \in \{1, \dots, L-1\}$, set the target accuracy $\varepsilon_\ell \coloneqq 64\kappa\sqrt{(\lceil \log_2(4k) \rceil + 1)mk \log(4mT/\delta)/B_\ell}$ and the sampling thresholds $\tau_{\ell,s} \coloneqq \lceil 64\kappa^2 s^2 \log(4mT/\delta)/\varepsilon_\ell^2 \rceil$ for $s \in \mathcal{S}_{\mathrm{dyad}}$.

Given a nonnegative integer pull-count vector $\mathbf{n} = (n_1, \dots, n_m)$, let $n_{\ell,s}(S; \mathbf{n}) = |\{i \in S : n_i < \tau_{\ell,s}\}|$ count the under-explored arms in S at scale s. For each feasible super arm $S \in \mathcal{I}$, define its dyadic deficiency score by

$$\mathrm{score}_\ell(S; \mathbf{n}) \coloneqq \sum_{s \in \mathcal{S}_{\mathrm{dyad}}} \frac{2}{s} n_{\ell,s}(S; \mathbf{n}). \tag{2}$$

The score measures the degree of under-exploration of a super arm across multiple scales. A super arm receives a larger score when it contains more base arms that remain under-explored. Since $n_{\ell,s}(S; \mathbf{n}) \leq |S| \leq k$ and $\sum_{s \in \mathcal{S}_{\mathrm{dyad}}} 2/s < 4$, we have $0 \leq \mathrm{score}_\ell(S; \mathbf{n}) < 4k$.

Algorithm description. Our algorithm maintains a current pull-count vector $\mathbf{N} = (N_1, \dots, N_m)$, initialized at $\mathbf{0}$. In each batch ℓ, it sequentially scans the active family $\mathcal{C}_{\ell-1}$. Whenever it encounters a super arm $S \in \mathcal{C}_{\ell-1}$ with $\mathrm{score}_\ell(S; \mathbf{N}) \geq 1$, it pulls S once and updates the corresponding base-arm counts. This process continues until either the batch budget is exhausted or every active super arm has score below 1. In the latter case, the remaining rounds of the batch are filled with arbitrary super arms from $\mathcal{C}_{\ell-1}$. The complete procedure is summarized in Algorithm 1.

5.2 Regret Upper Bound

Theorem 1 (High-probability pseudo-α-regret bound)**.** *Assume the stochastic CMAB model above and Assumptions 1–3. Algorithm 1 uses $L = \lceil \log_2 \log_2 T \rceil + 1$ batches. For every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$R_\alpha(T) \leq O\left(\sqrt{mT \log(2k)}\left(\Delta_{\max} + \kappa\sqrt{k \log(mT/\delta)} \log\log T\right)\right).$$

Corollary 1 (Expected regret)**.** *Under the conditions of Theorem 1, choosing $\delta = T^{-1}$ gives*

$$\mathbb{E}[R_\alpha(T)] \leq O\left(\sqrt{mT \log(2k)}\left(\Delta_{\max} + \kappa\sqrt{k \log(mT)} \log\log T\right)\right).$$

The bounds with explicit constants and their proofs are given in Appendix A.

Up to logarithmic factors and the initialization term, SE−BC achieves $\widetilde{O}(\kappa\sqrt{mkT})$ regret using only $O(\log\log T)$ batches. This matches the regret order of CUCB while reducing its batch complexity from $\Theta(T)$ to $O(\log\log T)$. For linear rewards with $\kappa = 1$, SE−BC also matches the $\widetilde{O}(\sqrt{mkT})$ regret of AROQ−CMAB with fewer batches, and improves over the $\widetilde{O}(k\sqrt{mT})$ regret of SROQ−CMAB while keeping the same $O(\log\log T)$ batch order.

Algorithm 1 SE−BC: **S**uccessive **E**limination for **B**atched **C**MAB

Require: $T \geq 4$, confidence parameter $\delta \in (0,1)$

Compute L, $(t_\ell)_{\ell=0}^{L}$, $(B_\ell)_{\ell=1}^{L}$, $\mathcal{S}_{\mathrm{dyad}}$, $(\varepsilon_\ell)_{\ell=1}^{L-1}$, and $(\tau_{\ell,s})_{\ell,s}$
Set $\mathcal{C}_0 \leftarrow \mathcal{I}$, $\Gamma_0 \leftarrow 0$, and initialize $\mathbf{N} \leftarrow \mathbf{0}$
for $\ell = 1, 2, \ldots, L-1$ **do**
  Fix an arbitrary order of the current active family $\mathcal{C}_{\ell-1}$
  $u \leftarrow 0$
  for each $S \in \mathcal{C}_{\ell-1}$ **do**
    while $u < B_\ell$ and $\mathrm{score}_\ell(S; \mathbf{N}) \geq 1$ **do** ▷ Score function is defined in (2)
      Play S once, update $\mathbf{N}$, and set $u \leftarrow u + 1$
    end while
  end for
  while $u < B_\ell$ **do**
    Play an arbitrary super arm in $\mathcal{C}_{\ell-1}$ once, update $\mathbf{N}$, and set $u \leftarrow u + 1$
  end while
  Set $\mathbf{N}_{t_\ell} \leftarrow \mathbf{N}$ and compute $\widehat{\mu}_{\ell,i}$, $\beta_{\ell,i}$, $\underline{\mu}_{\ell,i}$, $\overline{\mu}_{\ell,i}$, and the box $\mathcal{M}_\ell$
  Compute $\widehat{S}_\ell \in \mathrm{Oracle}_\alpha(\underline{\boldsymbol{\mu}}_\ell)$ and update $\Gamma_\ell \leftarrow \max\{\Gamma_{\ell-1}, r_{\underline{\boldsymbol{\mu}}_\ell}(\widehat{S}_\ell)\}$
  Set $\mathcal{C}_\ell \leftarrow \{S \in \mathcal{I} : r_{\overline{\boldsymbol{\mu}}_\ell}(S) \geq \Gamma_\ell\}$ ▷ Elimination
  if $\mathcal{C}_\ell = \varnothing$ **then**
    $\mathcal{C}_\ell \leftarrow \{\widehat{S}_\ell\}$
  end if
end for
In the final batch L, play an arbitrary super arm in $\mathcal{C}_{L-1}$ for all remaining B_L rounds

6 Lower Bound

We now prove that the $O(\log\log T)$ batch complexity of SE−BC is optimal up to constant factors. The lower bound is established on a linear-reward Bernoulli subclass of the CMAB model. The hard instance is a disjoint k-path construction, following the spirit of Kveton et al. (2015c), while the adaptive-grid argument follows the crossing-event template used for batched bandits (Gao et al., 2019). Since the subclass below satisfies Assumptions 1–2 with $\kappa = 1$, any lower bound for this subclass also applies to the full model. For the lower-bound analysis, we use the standard pseudo-regret $R(T) \coloneqq \sum_{t=1}^{T} \big(\mathrm{OPT} - r_{\boldsymbol{\mu}}(S_t)\big)$.

Assume that $m = dk$ for an integer $d \geq 2$. Partition the base arms into d disjoint path super arms $\mathcal{P}_j \coloneqq \{(j-1)k + 1, \ldots, jk\}, \quad j \in [d]$, and set $\mathcal{I} = \{\mathcal{P}_1, \ldots, \mathcal{P}_d\}$. For each path j, the selected coordinates share one Bernoulli observation: when $\mathcal{P}_j$ is played for the u-th time, the learner observes $Y_{j,u}$ on each of the k base arms in $\mathcal{P}_j$, where $Y_{j,u}$ has mean p_j and the variables are independent across j and u. This uses only contemporaneous correlation among base arms, which is allowed by our model. Under the linear reward, $r_{\boldsymbol{\mu}}(\mathcal{P}_j) = \sum_{i \in \mathcal{P}_j} \mu_i = kp_j$. We denote this Bernoulli k-path subclass by $\mathfrak{P}_{d,k}$. Regret on this subclass is exactly k times the regret of the underlying d-armed Bernoulli bandit.

Theorem 2 (Adaptive-grid lower bound for Bernoulli k-paths)**.** *Assume $d \leq T$. Let Π_M be the class of adaptive M-batch policies and $\gamma_M = (2 - 2^{1-M})^{-1}$. Then*

$$\inf_{\pi \in \Pi_M} \sup_{P \in \mathfrak{P}_{d,k}} \mathbb{E}_P^\pi[R(T)] = \Omega\left(M^{-2} k d^{1-\gamma_M} T^{\gamma_M}\right).$$

Here $\mathbb{E}_P^\pi$ denotes expectation under policy π on instance P, and the implicit constant is universal.

Corollary 2 (Batch-complexity barrier)**.** *Fix any constant $\zeta \in (0,1]$ and suppose $d \leq T^{1-\zeta}$. If a family of adaptive $M(T)$-batch policies satisfies*

$$\sup_{\textit{CMAB instances}} \mathbb{E}[R(T)] \leq \sqrt{mkT}\,\mathrm{polylog}(m,T)$$

for all sufficiently large T, then

$$M(T) = \Omega(\log\log T).$$

The proof of Theorem 2 and Corollary 2 is deferred to Appendix B.

Corollary 2 shows that, in the regime covered by the corollary, achieving near-optimal worst-case regret requires $\Omega(\log \log T)$ batches. Together with Corollary 1, this shows that SE−BC achieves near-optimal regret with the optimal order of batch complexity whenever the conditions of the two results are satisfied.

7 EXPERIMENTS

We compare SE−BC, CUCB (Chen et al., 2013), and AROQ (Kim et al., 2025) on two linear CMAB instances: Top-K with $m = 20$, $k = 5$, and $|\mathcal{I}| = 15504$; and maximum spanning trees of K_7 with $m = 21$, $k = 6$, and $|\mathcal{I}| = 16807$. Arm means are sampled independently from $\mathrm{Uniform}(0, 1)$ and observations are Bernoulli. We use exact Top-K and maximum spanning-tree oracles, $T = 10000$, $\alpha = 1$, and $\delta = 1/T$. Results are averaged over 100 independent runs, with the same mean vector shared across algorithms in each run. Table entries and shaded regions report sample standard deviations.

We follow the prescribed batch schedule and dyadic exploration sweep. After all active scores fall below 1, set $\widehat{\mu}_{0,i} = 0$, $g(n) := 2^{-\frac{1}{2}\lfloor \log_2 \max\{1,n\} \rfloor}$, and $\lambda_\ell := (L - \ell)/L$, and select $S \in \mathcal{C}_{\ell-1}$ maximizing

$$\frac{1}{k} \sum_{i \in S} \big(\widehat{\mu}_{\ell-1,i} + \lambda_\ell g(N_{t-1,i}) \big).$$

Here $\widehat{\mu}_{\ell-1,i}$ is the empirical mean after batch $\ell - 1$ and $N_{t-1,i}$ is the pull count before round t. Ties favor higher empirical reward. The empirical term favors actions with strong historical performance, whereas $g(N_{t-1,i})$ favors less-sampled arms. Since λ_ℓ decreases across batches, the allocation gradually shifts from exploration to exploitation.

We play the selected action until one of its g terms changes or the batch ends, then reselect if rounds remain. In the final batch, $\lambda_L = 0$ and the empirically best active action is played throughout. This rule requires no within-batch reward feedback and retains five batches. Reported oracle calls count benchmark updates and action reselection.

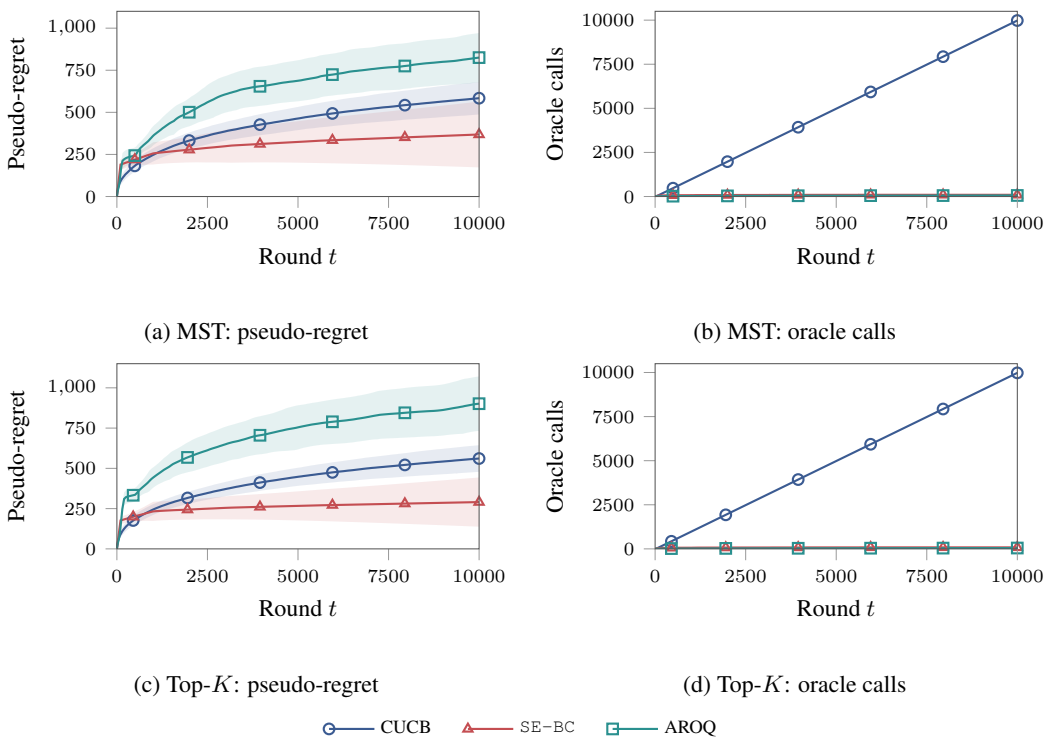
As summarized in Table 2 and Figure 1, SE−BC achieves the lowest cumulative pseudo-regret on both instances. This improvement mainly comes from its within-batch allocation strategy. SE−BC uses updated pull counts to prioritize under-sampled arms. After sufficient exploration, it shifts toward actions with higher empirical rewards. This better balances exploration and exploitation within each batch, explaining the lower regret observed in Figure 1. The final actions also achieve over 99.8% of the optimal expected reward. Meanwhile, its oracle cost remains competitive with the baselines.

Table 2: Statistical Comparison of Algorithm Performance

Instance	Algorithm	Regret	Oracle calls	Batches	$r_\mu(S_T)/\,\mathrm{OPT}$
MST	CUCB	584.03 ± 96.57	9979.00 ± 0.00	10000	0.9961 ± 0.0130
MST	AROQ	825.22 ± 146.12	$\mathbf{59.95 \pm 4.87}$	60.95	0.9944 ± 0.0177
MST	SE−BC	$\mathbf{368.50 \pm 195.48}$	97.17 ± 9.60	**5**	$\mathbf{0.9982 \pm 0.0039}$
Top-K	CUCB	561.04 ± 82.83	9980.00 ± 0.00	10000	0.9970 ± 0.0082
Top-K	AROQ	901.58 ± 168.76	$\mathbf{47.96 \pm 5.01}$	48.96	0.9947 ± 0.0173
Top-K	SE−BC	$\mathbf{290.59 \pm 152.79}$	85.78 ± 7.72	**5**	$\mathbf{0.9989 \pm 0.0035}$

8 CONCLUSION

In this work, we studied the batched combinatorial semi-bandit problem and proposed SE−BC. Under monotonicity and additive smoothness, SE−BC achieved near-optimal $\widetilde{O}(\kappa\sqrt{mkT})$ pseudo-α-regret using only $O(\log \log T)$ batches. We further established an $\Omega(\log \log T)$ lower bound on the number of batches required to achieve near-optimal worst-case regret, showing that SE−BC achieved the optimal order of batch complexity under the conditions of our results.



(a) MST: pseudo-regret

(b) MST: oracle calls

(c) Top-K: pseudo-regret

(d) Top-K: oracle calls

Figure 1: Mean cumulative pseudo-regret and oracle calls over 100 runs. Shading denotes one sample standard deviation.

Our analysis focused on worst-case regret. A natural direction for future work is to study problem-dependent regret for batched CMAB and investigate whether bounds of order $\widetilde{O}(km/\Delta_{\min})$ can be achieved with a small number of batches, where $\Delta_{\min}$ denotes the minimum suboptimality gap. Such bounds would provide a finer characterization of the performance of SE−BC on easier problem instances.

AI Use Statement

In this work, the research problem, core ideas, methodological framework, and experimental design were developed by the authors. We used generative AI tools to assist with mathematical derivations, the formulation and writing of proofs, and parts of code development and visualization. The main experimental implementation was carried out by the authors, with simulated data generated through random sampling.

We have reviewed all AI-assisted work. In particular, AI-assisted mathematical derivations and proofs were carefully verified and revised by the authors, and AI-assisted code and visualizations were also reviewed and validated.

We take responsibility for the final content of this work, including text, claims, proofs, code, or artifacts produced with the aid of generative AI.

References

Niklas Åkerblom, Yuxin Chen, and Morteza Haghir Chehreghani. Online learning of energy consumption for navigation of electric vehicles. *Artificial Intelligence*, 317:103879, 2023.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pp. 151–159. PMLR, 2013.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN)*, pp. 1–9. IEEE, 2010.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Tanmay Goyal, Sukruta Prakash Midigeshi, and Gaurav Sinha. Contextual slate glm bandits with limited adaptivity. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.31449. Accepted at ICML 2026; arXiv:2606.31449.

Osama Hanna, Lin Yang, and Christina Fragouli. Efficient batched algorithm for contextual linear bandits with large action space via soft elimination. *Advances in Neural Information Processing Systems*, 36:56772–56783, 2023.

Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Quanquan Gu. Double explore-then-commit: Asymptotic optimality and beyond. *arXiv preprint arXiv:2002.09174*, 2020.

Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanquan Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In *International Conference on Machine Learning*, pp. 5065–5073. PMLR, 2021.

Jung-hun Kim, Milan Vojnovic, and Min-hwan Oh. Oracle-efficient combinatorial semi-bandits. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 24981–25016. Curran Associates, Inc., 2025. doi: 10.52202/085713-0840. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/237196fdf3472b8de08bdd1103bb7b67-Paper-Conference.pdf.

Fang Kong, Yueran Yang, Wei Chen, and Shuai Li. The hardness analysis of thompson sampling for combinatorial semi-bandits with greedy oracle. *Advances in Neural Information Processing Systems*, 34:26701–26713, 2021.

Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 767–776, 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 28, 2015b.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–543. PMLR, 2015c.

Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1245–1253, 2016.

Xutong Liu, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John Lui, and Wei Chen. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. *Advances in Neural Information Processing Systems*, 35:14904–14916, 2022.

Nadav Merlis and Shie Mannor. Batch-size independent regret bounds for the combinatorial multi-armed bandit problem. In *Conference on Learning Theory*, pp. 2465–2489. PMLR, 2019.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, pp. 660–681, 2016.

Pierre Perrault, Etienne Boursier, Michal Valko, and Vianney Perchet. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *Advances in Neural Information Processing Systems*, 33:5429–5440, 2020.

Xuanfei Ren, Tianyuan Jin, and Pan Xu. Optimal batched linear bandits. *arXiv preprint arXiv:2406.04137*, 2024.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/a8e864d04c95572d1aece099af852d0a-Abstract.html.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1113–1122. PMLR, 2015. URL https://proceedings.mlr.press/v37/wen15.html.

Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. Online influence maximization under independent cascade model with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Raymond Zhang and Richard Combes. Thompson sampling for combinatorial bandits: Polynomial regret and mismatched sampling paradox. *Advances in Neural Information Processing Systems*, 37:89437–89467, 2024.

A Proofs for the upper bound

A.1 A standard concentration event

We establish a uniform confidence event for the observations collected under adaptive action selection. Write $X_{i,u}$ for the outcome observed on the u-th pull of base arm i. Only samples actually collected by time T are included in the event below. The proof applies the Hoeffding–Azuma inequality for bounded martingale differences under the conditional-mean assumption in the stochastic model.

Proposition 1 (Uniform confidence event)**.** *Define the event*

$$\mathcal{E} \coloneqq \left\{ \forall i \in [m], \quad \forall n \in \{1, \dots, N_{T,i}\} : \left| \frac{1}{n} \sum_{u=1}^{n} X_{i,u} - \mu_i \right| \le \sqrt{\frac{\log(4mT/\delta)}{2n}} \right\}.$$

Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$*. Moreover, on* $\mathcal{E}$*,*

$$\boldsymbol{\mu} \in \mathcal{M}_\ell, \quad \forall \ell \in \{0, 1, \dots, L-1\}.$$

Proof. Fix an arm $i \in [m]$ and a sample size $n \in [T]$. For the analysis only, extend the observed sequence $(X_{i,u})$ beyond $N_{T,i}$ with independent $[0,1]$-valued variables of mean μ_i. Because each pull is selected before its outcome is observed, the conditional-mean assumption ensures that the centered observations, including this auxiliary extension, form martingale differences with respect to the histories at successive observation times. Each difference lies in $[-\mu_i, 1-\mu_i]$, an interval of length one. The Hoeffding–Azuma inequality therefore gives

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{u=1}^{n} X_{i,u} - \mu_i \right| > \sqrt{\frac{\log(4mT/\delta)}{2n}} \right) \le \frac{\delta}{2mT}.$$

A union bound over $i \in [m]$ and $n \in [T]$ yields $\mathbb{P}(\mathcal{E}^c) \le \delta/2 \le \delta$, since $\mathcal{E}$ concerns only the samples actually collected.

Now fix $\ell \in \{1, \dots, L-1\}$ and $i \in [m]$. If $N_{t_\ell,i} = 0$, then $\widehat{\mu}_{\ell,i} = 0$ and $\beta_{\ell,i} = 1$, so the raw confidence interval contains $\mu_i \in [0,1]$. If $N_{t_\ell,i} = n \ge 1$, then $n \le N_{T,i}$, and on $\mathcal{E}$,

$$|\widehat{\mu}_{\ell,i} - \mu_i| \le \sqrt{\frac{\log(4mT/\delta)}{2n}} = \beta_{\ell,i}.$$

Thus every raw interval contains μ_i. Intersecting these intervals recursively with the initial interval $[0,1]$ preserves this inclusion, giving

$$\underline{\mu}_{\ell,i} \le \mu_i \le \overline{\mu}_{\ell,i}, \quad \forall i \in [m].$$

Hence $\boldsymbol{\mu} \in \mathcal{M}_\ell$ for every nonterminal batch. The case $\ell = 0$ follows from $\mathcal{M}_0 = [0,1]^m$. □

A.2 Certified benchmark and active-family properties

Proposition 2 (Nested confidence boxes)**.** *For every $\ell \in \{1, \dots, L-1\}$,*

$$\mathcal{M}_\ell \subseteq \mathcal{M}_{\ell-1}.$$

Proof. Fix $\ell \in \{1, \dots, L-1\}$ and let $\boldsymbol{\nu} \in \mathcal{M}_\ell$. By definition,

$$\underline{\mu}_{\ell,i} \le \nu_i \le \overline{\mu}_{\ell,i}, \quad \forall i \in [m].$$

Because

$$\underline{\mu}_{\ell,i} = \max\{\underline{\mu}_{\ell-1,i}, \widehat{\mu}_{\ell,i} - \beta_{\ell,i}\} \ge \underline{\mu}_{\ell-1,i}$$

and

$$\overline{\mu}_{\ell,i} = \min\{\overline{\mu}_{\ell-1,i}, \widehat{\mu}_{\ell,i} + \beta_{\ell,i}\} \le \overline{\mu}_{\ell-1,i},$$

we obtain

$$\underline{\mu}_{\ell-1,i} \le \nu_i \le \overline{\mu}_{\ell-1,i}, \quad \forall i \in [m].$$

Hence $\boldsymbol{\nu} \in \mathcal{M}_{\ell-1}$. □

Proposition 3 (Benchmark validity, activity of $S^\star$, and shrinking active families)**.** *On the event $\mathcal{E}$, the following statements hold for every $\ell \in \{0, 1, \dots, L-1\}$.*

(i) $\Gamma_\ell \le \mathrm{OPT}$.

(ii) *If $\ell \ge 1$, then* $\Gamma_\ell \ge \alpha \max_{A \in \mathcal{I}} r_{\underline{\boldsymbol{\mu}}_\ell}(A)$.

(iii) $S^\star \in \mathcal{C}_\ell$.

(iv) *If $\ell \ge 1$, then* $\mathcal{C}_\ell \subseteq \mathcal{C}_{\ell-1}$.

Proof. We first prove (i) by induction on ℓ. Since $\mathrm{OPT} \ge 0$ by assumption, the case $\ell = 0$ follows from $\Gamma_0 = 0 \le \mathrm{OPT}$. Now fix $\ell \ge 1$ and assume $\Gamma_{\ell-1} \le \mathrm{OPT}$. On $\mathcal{E}$, Proposition 1 gives $\underline{\boldsymbol{\mu}}_\ell \le \boldsymbol{\mu}$ coordinatewise. By monotonicity,

$$r_{\underline{\boldsymbol{\mu}}_\ell}(\widehat{S}_\ell) \le \max_{A \in \mathcal{I}} r_{\underline{\boldsymbol{\mu}}_\ell}(A) \le \max_{A \in \mathcal{I}} r_{\boldsymbol{\mu}}(A) = \mathrm{OPT}.$$

Therefore

$$\Gamma_\ell = \max\{\Gamma_{\ell-1}, r_{\underline{\boldsymbol{\mu}}_\ell}(\widehat{S}_\ell)\} \le \mathrm{OPT},$$

which proves (i).

For (ii), if $\ell \ge 1$ then by definition of $\widehat{S}_\ell$,

$$r_{\underline{\boldsymbol{\mu}}_\ell}(\widehat{S}_\ell) \ge \alpha \max_{A \in \mathcal{I}} r_{\underline{\boldsymbol{\mu}}_\ell}(A).$$

Since $\Gamma_\ell \ge r_{\underline{\boldsymbol{\mu}}_\ell}(\widehat{S}_\ell)$, we obtain (ii).

For (iii), combine (i) with monotonicity and $\overline{\boldsymbol{\mu}}_\ell \geq \boldsymbol{\mu}$ on $\mathcal{E}$:

$$r_{\overline{\boldsymbol{\mu}}_\ell}(S^\star) \geq r_{\boldsymbol{\mu}}(S^\star) = \mathrm{OPT} \geq \Gamma_\ell.$$

Hence $S^\star \in \mathcal{C}_\ell$.

For (iv), fix $\ell \geq 1$ and take $S \in \mathcal{C}_\ell$. Then

$$r_{\overline{\boldsymbol{\mu}}_\ell}(S) \geq \Gamma_\ell.$$

By construction, $\Gamma_\ell \geq \Gamma_{\ell-1}$. Also $\overline{\boldsymbol{\mu}}_\ell \leq \overline{\boldsymbol{\mu}}_{\ell-1}$ coordinatewise, so monotonicity gives

$$r_{\overline{\boldsymbol{\mu}}_{\ell-1}}(S) \geq r_{\overline{\boldsymbol{\mu}}_\ell}(S) \geq \Gamma_\ell \geq \Gamma_{\ell-1}.$$

Thus $S \in \mathcal{C}_{\ell-1}$. □

A.3 From dyadic scores to pseudo-α-gaps

Lemma 1 (Successor-scale bound)**.** *Fix a nonterminal batch index $\ell \in \{1, \ldots, L-1\}$ and a count vector $\mathbf{n} \in \mathbb{N}^m$. For each arm $i \in [m]$, define*

$$s_i(\mathbf{n}) \coloneqq \max\{s \in \mathcal{S}_{\mathrm{dyad}} : n_i \geq \tau_{\ell,s}\},$$

with the convention $s_i(\mathbf{n}) = 0$ if the set is empty. Then for every feasible super arm $S \in \mathcal{I}$,

$$\sum_{i \in S} \frac{1}{s_i(\mathbf{n}) \vee 1} \leq \mathrm{score}_\ell(S; \mathbf{n}) + \frac{1}{4}.$$

In particular, if $\mathrm{score}_\ell(S; \mathbf{n}) < 1$, then $s_i(\mathbf{n}) \geq 1$ for every $i \in S$ and

$$\sum_{i \in S} \frac{1}{s_i(\mathbf{n})} < \frac{5}{4}.$$

Proof. If some arm $i \in S$ satisfies $s_i(\mathbf{n}) = 0$, then $n_i < \tau_{\ell,1}$, so $i \in \{j \in [m] : n_j < \tau_{\ell,1}\}$ and therefore

$$\mathrm{score}_\ell(S; \mathbf{n}) \geq \frac{2}{1} |\{i \in S : n_i < \tau_{\ell,1}\}| \geq 2.$$

This immediately proves the *in particular* statement when $\mathrm{score}_\ell(S; \mathbf{n}) < 1$. It remains to establish the general inequality.

Fix $i \in S$. If $s_i(\mathbf{n}) = 0$, then $1/(s_i(\mathbf{n}) \vee 1) = 1 \leq \mathrm{score}_\ell(S; \mathbf{n})$, so there is nothing to prove. Suppose now that $s_i(\mathbf{n}) \geq 1$.

If $s_i(\mathbf{n}) < 2^q$, then $2s_i(\mathbf{n}) \in \mathcal{S}_{\mathrm{dyad}}$. By maximality of $s_i(\mathbf{n})$,

$$n_i < \tau_{\ell, 2s_i(\mathbf{n})},$$

so $i \in \{j \in [m] : n_j < \tau_{\ell, 2s_i(\mathbf{n})}\}$. Hence the single term corresponding to $2s_i(\mathbf{n})$ in the score contributes

$$\frac{2}{2s_i(\mathbf{n})} = \frac{1}{s_i(\mathbf{n})}$$

for arm i.

If $s_i(\mathbf{n}) = 2^q$, then

$$\frac{1}{s_i(\mathbf{n})} = \frac{1}{2^q}.$$

Summing this bound over all such arms gives at most

$$\frac{|S|}{2^q} \leq \frac{k}{2^q} \leq \frac{1}{4},$$

because $2^q \geq 4k$.

Adding the two cases over all $i \in S$ yields

$$\sum_{i \in S} \frac{1}{s_i(\mathbf{n}) \vee 1} \leq \mathrm{score}_\ell(S; \mathbf{n}) + \frac{1}{4}.$$

The final statement follows when $\mathrm{score}_\ell(S; \mathbf{n}) < 1$. □

Proposition 4 (Gap bound from the revised active family)**.** *Assume that $\mathcal{E}$ holds. Fix a nonterminal batch index $\ell \in \{1, \dots, L-1\}$ and a super arm $S \in \mathcal{C}_\ell$. Suppose that*

$$\mathrm{score}_\ell(A; \mathbf{N}_{t_\ell}) < 1, \quad \forall A \in \mathcal{C}_{\ell-1}.$$

Then

$$\alpha\mathrm{OPT} - r_{\boldsymbol{\mu}}(S) \le \varepsilon_\ell.$$

Proof. Because $S \in \mathcal{C}_\ell$, the definition of the active family gives

$$r_{\overline{\boldsymbol{\mu}}_\ell}(S) \ge \Gamma_\ell.$$

By Proposition 3(ii),

$$\Gamma_\ell \ge \alpha r_{\underline{\boldsymbol{\mu}}_\ell}(S^\star).$$

Hence

$$\begin{aligned}\alpha\mathrm{OPT} - r_{\boldsymbol{\mu}}(S) &= \alpha r_{\boldsymbol{\mu}}(S^\star) - r_{\boldsymbol{\mu}}(S) \\ &\le \alpha\big(r_{\boldsymbol{\mu}}(S^\star) - r_{\underline{\boldsymbol{\mu}}_\ell}(S^\star)\big) + \big(r_{\overline{\boldsymbol{\mu}}_\ell}(S) - r_{\boldsymbol{\mu}}(S)\big).\end{aligned}$$

Indeed, after adding and subtracting $\alpha r_{\underline{\boldsymbol{\mu}}_\ell}(S^\star)$ and Γ_ℓ, the middle terms are nonpositive because $\alpha r_{\underline{\boldsymbol{\mu}}_\ell}(S^\star) \le \Gamma_\ell \le r_{\overline{\boldsymbol{\mu}}_\ell}(S)$.

We next bound each difference on the right-hand side. Proposition 3(iii)–(iv) gives

$$S^\star \in \mathcal{C}_\ell \subseteq \mathcal{C}_{\ell-1} \quad \text{and} \quad S \in \mathcal{C}_\ell \subseteq \mathcal{C}_{\ell-1}.$$

Therefore the hypothesis of the proposition implies

$$\mathrm{score}_\ell(S^\star; \mathbf{N}_{t_\ell}) < 1 \quad \text{and} \quad \mathrm{score}_\ell(S; \mathbf{N}_{t_\ell}) < 1.$$

Fix a generic super arm $A \in \{S, S^\star\}$. For each arm $i \in A$, define $s_i \coloneqq s_i(\mathbf{N}_{t_\ell})$ as in Lemma 1. Then Lemma 1 gives

$$s_i \ge 1 \text{for all } i \in A, \quad \sum_{i \in A} \frac{1}{s_i} < \frac{5}{4}.$$

Now $N_{t_\ell, i} \ge \tau_{\ell, s_i}$ by definition of s_i, so

$$\beta_{\ell,i} = \sqrt{\frac{\log(4mT/\delta)}{2N_{t_\ell,i}}} \le \sqrt{\frac{\log(4mT/\delta)}{2\tau_{\ell,s_i}}} \le \frac{\varepsilon_\ell}{\sqrt{128}\kappa s_i} < \frac{\varepsilon_\ell}{10\kappa s_i}.$$

On $\mathcal{E}$, the true mean and the confidence envelopes satisfy $\underline{\boldsymbol{\mu}}_\ell \le \boldsymbol{\mu} \le \overline{\boldsymbol{\mu}}_\ell$. Since $\underline{\mu}_{\ell,i} \ge \widehat{\mu}_{\ell,i} - \beta_{\ell,i}$, $\overline{\mu}_{\ell,i} \le \widehat{\mu}_{\ell,i} + \beta_{\ell,i}$, and $|\widehat{\mu}_{\ell,i} - \mu_i| \le \beta_{\ell,i}$, we obtain

$$0 \le \mu_i - \underline{\mu}_{\ell,i} \le 2\beta_{\ell,i} \quad \text{and} \quad 0 \le \overline{\mu}_{\ell,i} - \mu_i \le 2\beta_{\ell,i}.$$

Summing over $i \in A$ yields

$$\sum_{i \in A}(\mu_i - \underline{\mu}_{\ell,i}) \le 2\sum_{i \in A}\beta_{\ell,i} < \frac{\varepsilon_\ell}{5\kappa}\sum_{i \in A}\frac{1}{s_i} < \frac{\varepsilon_\ell}{4\kappa},$$

and likewise

$$\sum_{i \in A}(\overline{\mu}_{\ell,i} - \mu_i) < \frac{\varepsilon_\ell}{4\kappa}.$$

Applying additive smoothness to $A = S^\star$ and $A = S$ gives

$$0 \le r_{\boldsymbol{\mu}}(S^\star) - r_{\underline{\boldsymbol{\mu}}_\ell}(S^\star) < \frac{\varepsilon_\ell}{4} \quad \text{and} \quad 0 \le r_{\overline{\boldsymbol{\mu}}_\ell}(S) - r_{\boldsymbol{\mu}}(S) < \frac{\varepsilon_\ell}{4}.$$

Substituting into the initial gap decomposition and using $\alpha \le 1$, we conclude that

$$\alpha\mathrm{OPT} - r_{\boldsymbol{\mu}}(S) < \frac{\varepsilon_\ell}{4} + \frac{\varepsilon_\ell}{4} = \frac{\varepsilon_\ell}{2} \le \varepsilon_\ell.$$

□

A.4 CHARGING THE DYADIC SWEEPS

Proposition 5 (Sweep termination inside a long enough batch)**.** *Fix $\ell \in \{1, \dots, L-1\}$ and assume that*

$$B_\ell \geq 8\big(\lceil \log_2(4k) \rceil + 1\big)m.$$

Then the dyadic sweep of Algorithm 1 terminates before the batch budget is exhausted, and the final counts satisfy

$$\mathrm{score}_\ell(S; \mathbf{N}_{t_\ell}) < 1, \quad \forall S \in \mathcal{C}_{\ell-1}.$$

Proof. Consider only the rounds inside batch ℓ that are triggered by the condition $\mathrm{score}_\ell(S; \mathbf{N}) \geq 1$. A round is called *triggered* if the algorithm plays a super arm because its current score is at least 1. All filler rounds are ignored in the charging argument.

For each triggered round t in batch ℓ, let $S(t)$ be the super arm played at that round. Because the round is triggered,

$$\mathrm{score}_\ell\big(S(t); \mathbf{N}_{t-1}\big) \geq 1.$$

Since there are exactly $\lceil \log_2(4k) \rceil + 1$ dyadic scales, there exists some $s(t) \in \mathcal{S}_{\mathrm{dyad}}$ such that

$$\frac{2}{s(t)} n_{\ell,s(t)}\big(S(t); \mathbf{N}_{t-1}\big) \geq \frac{1}{\big(\lceil \log_2(4k) \rceil + 1\big)}.$$

Equivalently,

$$n_{\ell,s(t)}\big(S(t); \mathbf{N}_{t-1}\big) \geq \frac{s(t)}{2\big(\lceil \log_2(4k) \rceil + 1\big)}. \tag{3}$$

Charge round t to the dyadic scale $s(t)$.

Fix a scale $s \in \mathcal{S}_{\mathrm{dyad}}$. Let $T_{\ell,s}$ be the set of triggered rounds in batch ℓ that are charged to s. By (3), every round $t \in T_{\ell,s}$ covers at least $s/(2\big(\lceil \log_2(4k) \rceil + 1\big))$ arms in the set

$$\{i \in [m] : N_{t-1,i} < \tau_{\ell,s}\}.$$

Hence every round in $T_{\ell,s}$ contributes at least $s/(2\big(\lceil \log_2(4k) \rceil + 1\big))$ incidences to the collection of pairs (t, i) such that

$$i \in S(t) \quad \text{and} \quad N_{t-1,i} < \tau_{\ell,s}.$$

On the other hand, for each fixed arm i, the indicator

$$\mathbb{1}\{N_{t-1,i} < \tau_{\ell,s}, i \in S(t)\}$$

can equal 1 at most $\tau_{\ell,s}$ times over the whole batch. Indeed, each such occurrence increases the count of arm i by one, and after $\tau_{\ell,s}$ such occurrences the arm permanently leaves the set $\{j : N_{t-1,j} < \tau_{\ell,s}\}$. Summing over all arms gives

$$\sum_{t \in T_{\ell,s}} \big|S(t) \cap \{i : N_{t-1,i} < \tau_{\ell,s}\}\big| \leq m\tau_{\ell,s}.$$

Combining this with the lower bound $s/(2\big(\lceil \log_2(4k) \rceil + 1\big))$ per charged round yields

$$|T_{\ell,s}| \leq \frac{2\big(\lceil \log_2(4k) \rceil + 1\big)}{s} m\tau_{\ell,s}. \tag{4}$$

Summing over all scales and using the definitions in Section 5.1,

$$\begin{aligned}
\#\{\text{triggered rounds in batch } \ell\} &\leq \sum_{s \in \mathcal{S}_{\mathrm{dyad}}} |T_{\ell,s}| \\
&\leq 2\big(\lceil \log_2(4k) \rceil + 1\big)m \sum_{s \in \mathcal{S}_{\mathrm{dyad}}} \frac{\tau_{\ell,s}}{s} \\
&\leq 2\big(\lceil \log_2(4k) \rceil + 1\big)m \sum_{s \in \mathcal{S}_{\mathrm{dyad}}} \left(\frac{1}{s} + \frac{B_\ell s}{64\big(\lceil \log_2(4k) \rceil + 1\big)mk} \right).
\end{aligned}$$

By $q = \lceil \log_2(4k) \rceil$, we have $2^q < 8k$ and $\sum_{s \in \mathcal{S}_{\mathrm{dyad}}} s = 2^{q+1} - 1 < 16k$. Together with $\sum_{s \in \mathcal{S}_{\mathrm{dyad}}} 1/s < 2$, this gives

$$\#\{\text{triggered rounds in batch } \ell\} < 4\big(\lceil \log_2(4k) \rceil + 1\big)m + \frac{B_\ell}{32k} \cdot 16k = 4\big(\lceil \log_2(4k) \rceil + 1\big)m + \frac{B_\ell}{2}.$$

If $B_\ell \geq 8\big(\lceil \log_2(4k) \rceil + 1\big)m$, then $4\big(\lceil \log_2(4k) \rceil + 1\big)m + B_\ell/2 \leq B_\ell$. The strict inequality above therefore bounds the number of triggered rounds by a quantity strictly smaller than B_ℓ, so the sweep finishes before the budget is exhausted. Once an action has been processed, its score remains below 1 because all counts are nondecreasing. Hence every action in $\mathcal{C}_{\ell-1}$ has score below 1 at the end of the sweep, and subsequent filler rounds preserve this property. Thus

$$\mathrm{score}_\ell(S; \mathbf{N}_{t_\ell}) < 1, \quad \forall S \in \mathcal{C}_{\ell-1}.$$

□

A.5 Schedule properties and proof of the main theorem

Proposition 6 (Elementary properties of the explicit schedule)**.** *Assume $T \geq 4$ and use the schedule defined in Section 5.1. Then the following statements hold.*

(i) *The first scheduled endpoint satisfies $t_1 \leq \sqrt{T}$.*

(ii) *For every $\ell \in \{1, 2, \ldots, L-1\}$,*
$$B_{\ell+1} \leq 2\sqrt{T t_\ell}.$$

(iii) *For every $\ell \in \{1, 2, \ldots, L-1\}$,*
$$t_\ell \leq 6B_\ell.$$

(iv) *The penultimate endpoint satisfies*
$$t_{L-1} \geq T/4.$$

(v) *The number of scheduled batches is exactly $L = \lceil \log_2 \log_2 T \rceil + 1$.*

Proof. Part (i) is immediate from
$$t_1 = \left\lfloor T^{1/2} \right\rfloor \leq \sqrt{T}.$$

For part (ii), first take $\ell \leq L-2$. Since $B_{\ell+1} = t_{\ell+1} - t_\ell \leq t_{\ell+1}$,
$$B_{\ell+1} \leq T^{1-2^{-(\ell+1)}} = \sqrt{T T^{1-2^{-\ell}}} < \sqrt{T(t_\ell + 1)} \leq 2\sqrt{T t_\ell},$$
because $t_\ell \geq 1$. For the final case $\ell = L-1$, one has $B_L \leq T$. Also,
$$2^{-(L-1)} \leq \frac{1}{\log_2 T},$$
so
$$t_{L-1} = \left\lfloor T^{1-2^{-(L-1)}} \right\rfloor \geq \left\lfloor T^{1-1/\log_2 T} \right\rfloor = \lfloor T/2 \rfloor \geq T/4,$$
because $T \geq 4$. Therefore
$$2\sqrt{T t_{L-1}} \geq 2\sqrt{T \cdot T/4} = T \geq B_L.$$
This proves part (ii).

For part (iii), when $\ell = 1$, $t_1 = B_1 \leq 6B_1$. Now fix $\ell \in \{2, \ldots, L-1\}$. Since $\ell \leq L - 1 = \lceil \log_2 \log_2 T \rceil$, we have
$$2^{-\ell} \geq 2^{-(L-1)} \geq \frac{1}{2 \log_2 T}.$$
Hence
$$T^{2^{-\ell}} \geq T^{1/(2\log_2 T)} = \sqrt{2}.$$

Using

$$t_{\ell-1} = \left\lfloor T^{1-2^{-(\ell-1)}} \right\rfloor = \left\lfloor T^{1-2^{-\ell}} / T^{2^{-\ell}} \right\rfloor,$$

we obtain

$$t_{\ell-1} \le \frac{T^{1-2^{-\ell}}}{\sqrt{2}} \le \frac{t_\ell + 1}{\sqrt{2}}.$$

Therefore

$$B_\ell = t_\ell - t_{\ell-1} \ge \left(1 - \frac{1}{\sqrt{2}}\right) t_\ell - \frac{1}{\sqrt{2}}.$$

The case $\ell \ge 2$ can occur only when $T \ge 5$, in which case $t_\ell \ge t_2 = \lfloor T^{3/4} \rfloor \ge 3$. The preceding lower bound is therefore strictly positive. Since B_ℓ is an integer, $B_\ell \ge 1$, and hence

$$t_\ell \le \frac{B_\ell + 1/\sqrt{2}}{1 - 1/\sqrt{2}} \le 6B_\ell.$$

The lower bound $t_{L-1} \ge T/4$ was already established in the final case of part (ii), which proves part (iv). Part (v) follows directly from the definition of L. □

Proof of Theorem 1. The batch-count statement is Proposition 6(v). We now prove the regret bound on the event $\mathcal{E}$ from Proposition 1; this event has probability at least $1 - \delta$.

Define the warm-up index

$$\ell_0 := \min\{\ell \in \{1, \dots, L-1\} : t_\ell \ge 48\big(\lceil \log_2(4k) \rceil + 1\big)m\},$$

with the convention $\ell_0 := L$ if the set is empty.

Warm-up rounds. Since $\mathrm{OPT} \ge 0$ and $\alpha \in (0, 1]$,

$$\alpha\mathrm{OPT} \le \mathrm{OPT} \le R_{\max}.$$

Together with $r_{\boldsymbol{\mu}}(S_t) \ge R_{\min}$, this bounds the pseudo-α-regret in every round t by

$$\alpha\mathrm{OPT} - r_{\boldsymbol{\mu}}(S_t) \le R_{\max} - R_{\min} = \Delta_{\max}.$$

If $\ell_0 = L$, then all T rounds are warm-up rounds. In this case $t_{L-1} < 48\big(\lceil \log_2(4k) \rceil + 1\big)m$, while Proposition 6(iv) gives $t_{L-1} \ge T/4$. Hence $T < 192\big(\lceil \log_2(4k) \rceil + 1\big)m$, so $\min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\} = T$. Consequently,

$$R_\alpha(T) \le \Delta_{\max} T = \Delta_{\max} \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\}.$$

Assume now that $\ell_0 \le L - 1$. If $\ell_0 = 1$, then

$$t_{\ell_0} = t_1 \le \sqrt{T} \le \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\}$$

by Proposition 6(i). If $\ell_0 \ge 2$, then $t_{\ell_0 - 1} < 48\big(\lceil \log_2(4k) \rceil + 1\big)m$ and Proposition 6(ii) gives

$$B_{\ell_0} \le 2\sqrt{T t_{\ell_0 - 1}} < 2\sqrt{48\big(\lceil \log_2(4k) \rceil + 1\big)mT} < 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}.$$

Hence

$$\begin{aligned} t_{\ell_0} &= t_{\ell_0 - 1} + B_{\ell_0} \\ &< 48\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT} \\ &\le \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\}. \end{aligned}$$

Therefore, in all cases,

$$t_{\ell_0} \le \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\}.$$

Since each of the first t_{ℓ_0} rounds incurs pseudo-α-regret at most $\Delta_{\max}$, we obtain

$$\sum_{t=1}^{t_{\ell_0}} \big(\alpha\mathrm{OPT} - r_{\boldsymbol{\mu}}(S_t)\big) \le \Delta_{\max} \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\}.$$

Later batches. Fix $\ell \in \{\ell_0, \ell_0 + 1, \dots, L-1\}$ and consider batch $\ell + 1$. Because $t_\ell \ge t_{\ell_0} \ge 48\big(\lceil \log_2(4k) \rceil + 1\big)m$, Proposition 6(iii) gives

$$B_\ell \ge t_\ell/6 \ge 8\big(\lceil \log_2(4k) \rceil + 1\big)m.$$

Hence Proposition 5 applies at the end of batch ℓ and yields

$$\mathrm{score}_\ell(A; \mathbf{N}_{t_\ell}) < 1, \quad \forall A \in \mathcal{C}_{\ell-1}.$$

Every super arm played in batch $\ell+1$ belongs to $\mathcal{C}_\ell$, and Proposition 3(iv) gives $\mathcal{C}_\ell \subseteq \mathcal{C}_{\ell-1}$. Therefore Proposition 4 applies to every super arm played in batch $\ell + 1$ and yields

$$\alpha\mathrm{OPT} - r_{\boldsymbol{\mu}}(S_t) \le \varepsilon_\ell, \quad \forall t \in \{t_\ell + 1, \dots, t_{\ell+1}\}.$$

Hence the regret of batch $\ell + 1$ is at most

$$B_{\ell+1}\varepsilon_\ell.$$

Using Proposition 6(ii) and the definitions in Section 5.1,

$$\begin{aligned}
B_{\ell+1}\varepsilon_\ell &\le 2\sqrt{Tt_\ell} \cdot 64\kappa \sqrt{\frac{\big(\lceil \log_2(4k) \rceil + 1\big)mk \log(4mT/\delta)}{B_\ell}} \\
&\le 2\sqrt{6TB_\ell} \cdot 64\kappa \sqrt{\frac{\big(\lceil \log_2(4k) \rceil + 1\big)mk \log(4mT/\delta)}{B_\ell}} \\
&= 128\sqrt{6}\kappa \sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mkT \log(4mT/\delta)} \\
&\le 320\kappa \sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mkT \log(4mT/\delta)}.
\end{aligned}$$

This bound is the same for every such ℓ. There are at most $L - 1$ later batches, so their total regret is at most

$$320\kappa \sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mkT \log(4mT/\delta)}(L-1).$$

Summing the batches. Combining the warm-up bound with the later-batch bound yields

$$\begin{aligned}
R_\alpha(T) \le \Delta_{\max} \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\} \\
+ 320\kappa \sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mkT \log(4mT/\delta)}(L-1)
\end{aligned}$$

on $\mathcal{E}$. Since $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, the claimed high-probability statement follows. The stated $O(\cdot)$ form uses $\lceil \log_2(4k) \rceil + 1 = O(\log k)$ and $L - 1 = O(\log \log T)$. □

Proof of Corollary 1. Apply Theorem 1 with $\delta = T^{-1}$. With probability at least $1 - T^{-1}$,

$$\begin{aligned}
R_\alpha(T) \le \Delta_{\max} \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\} \\
+ 320\kappa \sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mkT \log(4mT^2)}(L-1).
\end{aligned}$$

On the complementary event, every single-round pseudo-α-regret term is at most $\Delta_{\max}$, so always

$$R_\alpha(T) \le \Delta_{\max} T.$$

Taking expectations and splitting according to the good event gives

$$\begin{aligned}
\mathbb{E}[R_\alpha(T)] &\le \Delta_{\max} \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\} \\
&\quad + 320\kappa \sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mkT \log(4mT^2)}(L-1) + T^{-1} \cdot \Delta_{\max} T \\
&= \Delta_{\max} \min\{T, 192\big(\lceil \log_2(4k) \rceil + 1\big)m + 16\sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mT}\} \\
&\quad + 320\kappa \sqrt{\big(\lceil \log_2(4k) \rceil + 1\big)mkT \log(4mT^2)}(L-1) + \Delta_{\max}.
\end{aligned}$$

□

B PROOFS FOR THE LOWER BOUND

Throughout this appendix section, fix an arbitrary adaptive M-batch policy on the Bernoulli k-path subclass $\mathfrak{P}_{d,k}$. Let

$$0 = \sigma_0 < \sigma_1 < \cdots < \sigma_M = T$$

be its possibly random batch endpoints, and let $J_t \in [d]$ be the path index played at round t, so the actual super arm is $\mathcal{P}_{J_t}$. Let $\mathcal{G}_b$ be the sigma-field generated by the policy's internal randomness and by all information revealed by the end of batch b. By the batched protocol, both σ_b and the entire batch-b action plan

$$(J_t)_{\sigma_{b-1}<t\le\sigma_b}$$

are $\mathcal{G}_{b-1}$-measurable.

For each fixed time $\bar{t} \in [T]$, let $\mathcal{H}_{\bar{t}}$ be the fixed-time omniscient history generated by the policy's internal randomness, the selected path indices $J_1, \ldots, J_{\bar{t}}$, and the latent Bernoulli bits generated on the selected paths in rounds $1, \ldots, \bar{t}$, whether or not those bits have been revealed to the learner by time $\bar{t}$. On the event $\{\sigma_b \le \bar{t}\}$ we have $\mathcal{G}_b \subseteq \mathcal{H}_{\bar{t}}$. For any instance law $\mathbb{P}$, write $\mathbb{P}^{\bar{t}}$ for the law of $\mathcal{H}_{\bar{t}}$, and define

$$N_v^{\mathrm{path}}(\bar{t}) \coloneqq \sum_{t=1}^{\bar{t}} \mathbb{1}\{J_t = v\}.$$

B.1 INFORMATION-THEORETIC INGREDIENTS

Lemma 2 (Bernoulli KL upper bound)**.** *If $p, q \in [1/4, 3/4]$, then*

$$\mathrm{kl}(p, q) \le 4(p-q)^2,$$

where $\mathrm{kl}(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$. *In particular, for every $u \in [0, 1/12]$,*

$$\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+u\right) \le 4u^2, \quad \mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+3u\right) \le 36u^2.$$

Proof. Fix $q \in [1/4, 3/4]$ and set $f_q(p) = \mathrm{kl}(p, q)$. Then $f_q(q) = f_q'(q) = 0$ and

$$f_q''(x) = \frac{1}{x(1-x)}.$$

For $x \in [1/4, 3/4]$, we have $x(1-x) \ge 3/16$, and hence $f_q''(x) \le 16/3 < 8$. Taylor's theorem gives

$$\mathrm{kl}(p, q) = \frac{1}{2} f_q''(\xi)(p-q)^2 \le 4(p-q)^2$$

for some ξ between p and q. If $u \in [0, 1/12]$, then $1/2 + 3u \in [1/2, 3/4]$, so the two displayed special cases follow. □

Lemma 3 (TV–KL bound)**.** *For any probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same measurable space,*

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \le \sqrt{\mathrm{KL}(\mathbb{P}\|\mathbb{Q})}.$$

Proof. This is a weaker form of Pinsker's inequality. For completeness, assume $\mathbb{P} \ll \mathbb{Q}$ and write $r = d\mathbb{P}/d\mathbb{Q}$. The elementary inequality

$$x\log x - x + 1 \ge \frac{(x-1)^2}{2(x+1)}, \quad x > 0,$$

implies

$$\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \ge \frac{1}{2}\int \frac{(r-1)^2}{r+1} d\mathbb{Q}.$$

By Cauchy–Schwarz,

$$\begin{aligned}\mathrm{TV}(\mathbb{P},\mathbb{Q}) &= \frac{1}{2}\int |r-1| d\mathbb{Q} \\ &\le \frac{1}{2}\left(\int \frac{(r-1)^2}{r+1} d\mathbb{Q}\right)^{1/2}\left(\int (r+1) d\mathbb{Q}\right)^{1/2} \\ &\le \sqrt{\mathrm{KL}(\mathbb{P}\|\mathbb{Q})},\end{aligned}$$

where $\int (r+1)d\mathbb{Q} = 2$. If $\mathbb{P}$ is not absolutely continuous with respect to $\mathbb{Q}$, the right-hand side is infinite and the claim is trivial. □

Lemma 4 (Fixed-time KL accounting)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two Bernoulli k-path instances that are identical except on one path $v \in [d]$. Suppose that path v has Bernoulli mean p under $\mathbb{P}$ and mean q under $\mathbb{Q}$. Then, for every fixed time $\bar{t} \in [T]$,*

$$\mathrm{KL}(\mathbb{P}^{\bar{t}}\|\mathbb{Q}^{\bar{t}}) = \mathrm{kl}(p,q)\mathbb{E}_{\mathbb{P}}\left[N_v^{\mathrm{path}}(\bar{t})\right].$$

Proof. Let ξ denote all internal randomness of the policy, and let $Z_t \in \{0,1\}$ be the Bernoulli bit observed from the selected path at round t. The omniscient history $\mathcal{H}_{\bar{t}}$ is equivalent to the sequence

$$(\xi, J_1, Z_1, \ldots, J_{\bar{t}}, Z_{\bar{t}}).$$

The conditional action kernel is the same under the two instances, because the policy uses the same internal randomness and the same realized history. Thus, by the chain rule for KL divergence, only the conditional laws of the Bernoulli bits contribute:

$$\mathrm{KL}(\mathbb{P}^{\bar{t}}\|\mathbb{Q}^{\bar{t}}) = \sum_{t=1}^{\bar{t}} \mathbb{E}_{\mathbb{P}}\left[\mathrm{KL}\left(\mathbb{P}(Z_t \mid \xi, J_1, Z_1, \ldots, J_t)\|\mathbb{Q}(Z_t \mid \xi, J_1, Z_1, \ldots, J_t)\right)\right].$$

If $J_t \neq v$, the two conditional Bernoulli laws agree. If $J_t = v$, their KL divergence is $\mathrm{kl}(p,q)$. Therefore

$$\mathrm{KL}(\mathbb{P}^{\bar{t}}\|\mathbb{Q}^{\bar{t}}) = \sum_{t=1}^{\bar{t}} \mathbb{E}_{\mathbb{P}}[\mathbb{1}\{J_t = v\}]\mathrm{kl}(p,q) = \mathrm{kl}(p,q)\mathbb{E}_{\mathbb{P}}[N_v^{\mathrm{path}}(\bar{t})].$$

□

Lemma 5 (Star testing inequality on an event)**.** *Let $\widetilde{\mathbb{Q}}_1, \ldots, \widetilde{\mathbb{Q}}_d$ be finite measures on a common measurable space, and let Ψ be a measurable map taking values in $[d]$. Then, for every center $v \in [d]$,*

$$\frac{1}{d}\sum_{w=1}^{d} \widetilde{\mathbb{Q}}_w(\Psi \neq w) \ge \frac{1}{d}\sum_{w \neq v} \int \min\{d\widetilde{\mathbb{Q}}_v, d\widetilde{\mathbb{Q}}_w\}.$$

Proof. Let $\lambda = \widetilde{\mathbb{Q}}_1 + \cdots + \widetilde{\mathbb{Q}}_d$, and write $q_w = d\widetilde{\mathbb{Q}}_w/d\lambda$. Then

$$\frac{1}{d}\sum_{w=1}^{d} \widetilde{\mathbb{Q}}_w(\Psi \neq w) = \frac{1}{d}\int \left(\sum_{w=1}^{d} q_w - q_\Psi\right) d\lambda \ge \frac{1}{d}\int \left(\sum_{w=1}^{d} q_w - \max_{r\in[d]} q_r\right) d\lambda.$$

It remains to use the pointwise inequality

$$\sum_{w=1}^{d} x_w - \max_{r\in[d]} x_r \ge \sum_{w\neq v} \min\{x_v, x_w\}$$

for every nonnegative vector $(x_1, \ldots, x_d)$. By symmetry, take $v = 1$. If $x_1 = \max_r x_r$, the inequality is immediate. Otherwise, if some x_r with $r \ge 2$ is maximal, then

$$\sum_{w=1}^{d} x_w - \max_r x_r = x_1 + \sum_{w\neq 1,r} x_w \ge x_1 + \sum_{w\neq 1,r} \min\{x_1, x_w\} = \sum_{w=2}^{d} \min\{x_1, x_w\}.$$

□

Lemma 6 (Event-wise overlap bound)**.** *For any probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same measurable space and any event A,*

$$\int_A \min\{d\mathbb{P}, d\mathbb{Q}\} \geq \mathbb{P}(A) - \mathrm{TV}(\mathbb{P}, \mathbb{Q}).$$

Proof. Using $\min\{a, b\} = a - (a - b)_+$,

$$\int_A \min\{d\mathbb{P}, d\mathbb{Q}\} = \mathbb{P}(A) - \int_A (d\mathbb{P} - d\mathbb{Q})_+ \geq \mathbb{P}(A) - \mathrm{TV}(\mathbb{P}, \mathbb{Q}).$$

□

B.2 Crossing events and comparison grid

Lemma 7 (Comparison grid)**.** *Let*

$$a_0 \coloneqq 0, \quad a_j \coloneqq \frac{1 - 2^{-j}}{1 - 2^{-M}}, \quad j = 1, \ldots, M,$$

and define deterministic comparison times

$$g_0 \coloneqq d, \quad g_j \coloneqq \left\lfloor d^{1-a_j} T^{a_j} \right\rfloor, \quad j = 1, \ldots, M.$$

Then $g_M = T$, the sequence $(g_j)_{j=0}^M$ is nondecreasing, and for every $j \in [M]$,

$$g_j \sqrt{\frac{d}{g_{j-1}}} \geq \frac{1}{2} d^{1-\gamma_M} T^{\gamma_M}, \quad \gamma_M = \frac{1}{2 - 2^{1-M}}.$$

Proof. Since $a_M = 1$, $g_M = T$. Since $T \geq d$ and a_j is increasing, $d^{1-a_j} T^{a_j}$ is nondecreasing in j, and so is $(g_j)_{j=0}^M$.

Because $d^{1-a_j} T^{a_j} \geq d \geq 2$,

$$g_j \geq \frac{1}{2} d^{1-a_j} T^{a_j}, \quad g_{j-1} \leq d^{1-a_{j-1}} T^{a_{j-1}}.$$

Hence

$$\begin{aligned} g_j \sqrt{\frac{d}{g_{j-1}}} &\geq \frac{1}{2} d^{1-a_j} T^{a_j} \cdot d^{1/2} \cdot d^{-(1-a_{j-1})/2} T^{-a_{j-1}/2} \\ &= \frac{1}{2} d^{1-a_j+1/2-(1-a_{j-1})/2} T^{a_j - a_{j-1}/2}. \end{aligned}$$

A direct calculation gives

$$a_j - \frac{a_{j-1}}{2} = \gamma_M, \quad 1 - a_j + \frac{1}{2} - \frac{1 - a_{j-1}}{2} = 1 - \gamma_M.$$

Substitution proves the claim. □

For $j \in [M]$, define

$$\Delta_j \coloneqq \frac{1}{96M} \sqrt{\frac{d}{g_{j-1}}}.$$

Since $g_{j-1} \geq d$, we have $\Delta_j \leq 1/(96M) \leq 1/12$. Let $\mathbb{Q}$ denote the baseline instance in which every path has mean $1/2$. For every $j \in [M]$ and $v \in [d]$, let $\mathbb{P}_{j,v}$ denote the instance in which path v has mean $1/2 + \Delta_j$ and all other paths have mean $1/2$.

Define the crossing events

$$\mathcal{X}_j \coloneqq \{\sigma_{j-1} < g_{j-1}, \sigma_j \geq g_j\}, \quad j \in [M].$$

Lemma 8 (Crossing events are visible from the earlier fixed-time history)**.** *The events $\mathcal{X}_1, \dots, \mathcal{X}_M$ cover the sample space: at least one crossing event occurs on every sample path. They need not be disjoint. Moreover, for every $j \in [M]$,*

$$\mathcal{X}_j \in \mathcal{H}_{g_{j-1}},$$

and, on $\mathcal{X}_j$, every action J_t with $t \le g_j$ is $\mathcal{H}_{g_{j-1}}$-measurable.

Proof. Because $\sigma_0 = 0 < g_0 = d$ and $\sigma_M = g_M = T$, every sample path has a smallest index j satisfying $\sigma_j \ge g_j$. For this index, $\sigma_{j-1} < g_{j-1}$, so $\mathcal{X}_j$ occurs. Thus the crossing events cover the sample space, although other crossing events may also occur on the same path.

Fix j. On $\{\sigma_{j-1} \ge g_{j-1}\}$, $\mathcal{X}_j$ is false. On $\{\sigma_{j-1} < g_{j-1}\}$, the endpoint σ_j and the whole batch-j action plan are $\mathcal{G}_{j-1}$-measurable. Since $\mathcal{G}_{j-1} \subseteq \mathcal{H}_{g_{j-1}}$ on this event, $\{\sigma_j \ge g_j\}$ is $\mathcal{H}_{g_{j-1}}$-measurable there. Hence $\mathcal{X}_j \in \mathcal{H}_{g_{j-1}}$.

Now fix $t \le g_j$. On $\mathcal{X}_j$, if $t \le \sigma_{j-1}$, then J_t is already included in the fixed-time history $\mathcal{H}_{g_{j-1}}$. If $\sigma_{j-1} < t \le g_j$, then $t \le \sigma_j$ on $\mathcal{X}_j$, so J_t belongs to the batch-j action plan, which is $\mathcal{G}_{j-1}$-measurable and hence $\mathcal{H}_{g_{j-1}}$-measurable on $\mathcal{X}_j$. □

Lemma 9 (Some crossing event is likely in its own instance)**.** *Let*

$$\rho_j \coloneqq \frac{1}{d} \sum_{v=1}^{d} \mathbb{P}_{j,v}(\mathcal{X}_j), \quad j \in [M].$$

Then

$$\sum_{j=1}^{M} \rho_j \ge \frac{1}{2}.$$

Consequently, there exists $j_\star \in [M]$ such that $\rho_{j_\star} \ge 1/(2M)$.

Proof. By Lemma 8, $\mathcal{X}_j \in \mathcal{H}_{g_{j-1}}$. Therefore

$$|\mathbb{P}_{j,v}(\mathcal{X}_j) - \mathbb{Q}(\mathcal{X}_j)| \le \mathrm{TV}(\mathbb{P}_{j,v}^{g_{j-1}}, \mathbb{Q}^{g_{j-1}}).$$

Under $\mathbb{Q}$ and $\mathbb{P}_{j,v}$, only path v differs before time g_{j-1}. Lemmas 4 and 2 give

$$\mathrm{KL}(\mathbb{Q}^{g_{j-1}} \| \mathbb{P}_{j,v}^{g_{j-1}}) = \mathrm{kl}\left(\frac{1}{2}, \frac{1}{2} + \Delta_j\right) \mathbb{E}_{\mathbb{Q}}\left[N_v^{\mathrm{path}}(g_{j-1})\right] \le 4\Delta_j^2 \mathbb{E}_{\mathbb{Q}}\left[N_v^{\mathrm{path}}(g_{j-1})\right].$$

Averaging over v and using $\sum_{v=1}^{d} N_v^{\mathrm{path}}(g_{j-1}) = g_{j-1}$ almost surely,

$$\frac{1}{d} \sum_{v=1}^{d} \mathrm{KL}(\mathbb{Q}^{g_{j-1}} \| \mathbb{P}_{j,v}^{g_{j-1}}) \le 4\Delta_j^2 \frac{g_{j-1}}{d} = \frac{1}{2304M^2}.$$

By Lemma 3, symmetry of total variation, and Jensen's inequality,

$$\frac{1}{d} \sum_{v=1}^{d} \mathrm{TV}(\mathbb{P}_{j,v}^{g_{j-1}}, \mathbb{Q}^{g_{j-1}}) \le \frac{1}{48M}.$$

Thus

$$|\rho_j - \mathbb{Q}(\mathcal{X}_j)| \le \frac{1}{48M}.$$

Since the crossing events cover the sample space, $\sum_{j=1}^{M} \mathbb{Q}(\mathcal{X}_j) \ge 1$. Summing the preceding inequalities over j gives

$$\sum_{j=1}^{M} \rho_j \ge \sum_{j=1}^{M} \mathbb{Q}(\mathcal{X}_j) - \frac{1}{48} \ge \frac{47}{48} \ge \frac{1}{2}.$$

□

Lemma 10 (A large crossing probability forces regret)**.** *Fix $j \in [M]$. If $\rho_j \geq 1/(2M)$, then*

$$\sup_{\text{instances in } \mathfrak{P}_{d,k}} \mathbb{E}[R(T)] \geq \frac{1}{512} M^{-2} k g_j \sqrt{\frac{d}{g_{j-1}}}.$$

Proof. Fix j and abbreviate $\Delta = \Delta_j$. For each reference path $v \in [d]$, define d candidate instances $\mathbb{Q}_{v,w}$, indexed by their unique optimal path $w \in [d]$, as follows. Set $\mathbb{Q}_{v,v} = \mathbb{P}_{j,v}$. For $w \neq v$, let $\mathbb{Q}_{v,w}$ be the instance in which path v has mean $1/2 + \Delta$, path w has mean $1/2 + 3\Delta$, and all other paths have mean $1/2$. Since $\Delta \leq 1/12$, all path means lie in $[1/2, 3/4]$.

Under $\mathbb{Q}_{v,v}$, any pull of a path other than v incurs expected regret $k\Delta$. Under $\mathbb{Q}_{v,w}$ with $w \neq v$, the unique optimal path is w, and every pull of a path other than w incurs expected regret at least $k\Delta$. Therefore, for every $w \in [d]$,

$$\mathbb{E}_{\mathbb{Q}_{v,w}}[R(T)] \geq k\Delta \sum_{t=1}^{g_j} \mathbb{Q}_{v,w}(\mathcal{X}_j, J_t \neq w).$$

Averaging over w yields, for every fixed v,

$$\sup_{\text{instances in } \mathfrak{P}_{d,k}} \mathbb{E}[R(T)] \geq \frac{k\Delta}{d} \sum_{t=1}^{g_j} \sum_{w=1}^{d} \mathbb{Q}_{v,w}(\mathcal{X}_j, J_t \neq w). \tag{5}$$

Fix $t \leq g_j$. By Lemma 8, $\mathcal{X}_j$ and J_t on $\mathcal{X}_j$ are $\mathcal{H}_{g_{j-1}}$-measurable. Apply Lemma 5 to the finite measures

$$\widetilde{\mathbb{Q}}_w(\cdot) \coloneqq \mathbb{Q}_{v,w}^{g_{j-1}}(\cdot \cap \mathcal{X}_j), \quad w \in [d],$$

with decision rule $\Psi = J_t$ on $\mathcal{X}_j$ and center v. Then

$$\frac{1}{d} \sum_{w=1}^{d} \mathbb{Q}_{v,w}(\mathcal{X}_j, J_t \neq w) \geq \frac{1}{d} \sum_{w \neq v} \int_{\mathcal{X}_j} \min\{d\mathbb{Q}_{v,v}^{g_{j-1}}, d\mathbb{Q}_{v,w}^{g_{j-1}}\}.$$

Substituting into (5) gives

$$\sup_{\text{instances in } \mathfrak{P}_{d,k}} \mathbb{E}[R(T)] \geq \frac{k\Delta g_j}{d} \sum_{w \neq v} \int_{\mathcal{X}_j} \min\{d\mathbb{Q}_{v,v}^{g_{j-1}}, d\mathbb{Q}_{v,w}^{g_{j-1}}\}. \tag{6}$$

By Lemma 6,

$$\int_{\mathcal{X}_j} \min\{d\mathbb{Q}_{v,v}^{g_{j-1}}, d\mathbb{Q}_{v,w}^{g_{j-1}}\} \geq \mathbb{Q}_{v,v}(\mathcal{X}_j) - \mathrm{TV}(\mathbb{Q}_{v,v}^{g_{j-1}}, \mathbb{Q}_{v,w}^{g_{j-1}}).$$

Under $\mathbb{Q}_{v,v}$ and $\mathbb{Q}_{v,w}$, only path w differs before time g_{j-1}. Lemmas 4 and 2 imply

$$\mathrm{KL}(\mathbb{Q}_{v,v}^{g_{j-1}} \| \mathbb{Q}_{v,w}^{g_{j-1}}) = \mathrm{kl}\left(\frac{1}{2}, \frac{1}{2} + 3\Delta\right) \mathbb{E}_{\mathbb{Q}_{v,v}}\left[N_w^{\mathrm{path}}(g_{j-1})\right] \leq 36\Delta^2 \mathbb{E}_{\mathbb{P}_{j,v}}\left[N_w^{\mathrm{path}}(g_{j-1})\right].$$

Averaging over $w \neq v$ and using $\sum_{w \neq v} N_w^{\mathrm{path}}(g_{j-1}) \leq g_{j-1}$,

$$\frac{1}{d-1} \sum_{w \neq v} \mathrm{KL}(\mathbb{Q}_{v,v}^{g_{j-1}} \| \mathbb{Q}_{v,w}^{g_{j-1}}) \leq 36\Delta^2 \frac{g_{j-1}}{d-1} = \frac{d}{256M^2(d-1)} \leq \frac{1}{128M^2},$$

where $d \geq 2$ is used in the last step. Lemma 3 and Jensen's inequality give

$$\frac{1}{d-1} \sum_{w \neq v} \mathrm{TV}(\mathbb{Q}_{v,v}^{g_{j-1}}, \mathbb{Q}_{v,w}^{g_{j-1}}) \leq \frac{1}{8M}.$$

Returning to (6),

$$\sup_{\text{instances in } \mathfrak{P}_{d,k}} \mathbb{E}[R(T)] \geq \frac{d-1}{d} k\Delta g_j \left(\mathbb{P}_{j,v}(\mathcal{X}_j) - \frac{1}{8M}\right).$$

This holds for every v, so averaging over v and using the definition of ρ_j gives

$$\sup_{\text{instances in } \mathfrak{P}_{d,k}} \mathbb{E}[R(T)] \geq \frac{d-1}{d} k\Delta g_j \left(\rho_j - \frac{1}{8M}\right).$$

If $\rho_j \geq 1/(2M)$, then $\rho_j - 1/(8M) \geq 3/(8M)$, and $(d-1)/d \geq 1/2$. Therefore

$$\sup_{\text{instances in } \mathfrak{P}_{d,k}} \mathbb{E}[R(T)] \geq \frac{1}{2} \cdot \frac{3}{8M} k\Delta g_j \geq \frac{1}{512} M^{-2} k g_j \sqrt{\frac{d}{g_{j-1}}},$$

where the final inequality absorbs the numerical constant from $\Delta = (96M)^{-1}\sqrt{d/g_{j-1}}$. □

Proof of Theorem 2. By Lemma 9, there exists $j_\star \in [M]$ such that $\rho_{j_\star} \geq 1/(2M)$. Lemma 10 then yields

$$\sup_{\text{instances in } \mathfrak{P}_{d,k}} \mathbb{E}[R(T)] \geq \frac{1}{512} M^{-2} k g_{j_\star} \sqrt{\frac{d}{g_{j_\star - 1}}}.$$

Lemma 7 gives

$$g_{j_\star} \sqrt{\frac{d}{g_{j_\star - 1}}} \geq \frac{1}{2} d^{1-\gamma_M} T^{\gamma_M}.$$

Absorbing the constant $1/1024$ into a universal constant c proves the first display. The equivalent form follows from $d = m/k$:

$$k d^{1-\gamma_M} T^{\gamma_M} = k \left(\frac{m}{k}\right)^{1-\gamma_M} T^{\gamma_M} = m^{1-\gamma_M} k^{\gamma_M} T^{\gamma_M}.$$

□

Proof of Corollary 2. By Theorem 2, every adaptive M-batch policy satisfies

$$\sup \mathbb{E}[R(T)] \geq c M^{-2} k d^{1-\gamma_M} T^{\gamma_M} = c M^{-2} k \sqrt{dT} \left(\frac{T}{d}\right)^{\gamma_M - 1/2}.$$

Since $m = dk$, we have $k\sqrt{dT} = \sqrt{mkT}$. Also,

$$\gamma_M - \frac{1}{2} = \frac{1}{2^{M+1} - 2}.$$

If $d \leq T^{1-\zeta}$, then $T/d \geq T^{\zeta}$, so

$$\sup \mathbb{E}[R(T)] \geq c M^{-2} \sqrt{mkT} T^{\zeta/(2^{M+1}-2)}.$$

Therefore an upper bound of order $\sqrt{mkT}\,\mathrm{polylog}(m,T)$ for all sufficiently large T would require

$$T^{\zeta/(2^{M+1}-2)} \leq C M^2 \mathrm{polylog}(m,T)$$

for some constant C. Taking logarithms gives

$$\frac{\log T}{2^M} \lesssim \log\log T,$$

which is possible only when $M = \Omega(\log\log T)$. □

C Example of Implementation of the Elimination Step

We explain how the active family in (1) can be handled efficiently when the approximation oracle is implemented by the greedy procedure of Kong et al. (2021). The key observation is that we never need to explicitly enumerate the super arms in $\mathcal{C}_\ell$. Instead, the condition

$$r_{\overline{\boldsymbol{\mu}}_\ell}(S) \geq \Gamma_\ell$$

can be checked during the sequential construction of a super arm.

Recall that the greedy oracle constructs a super arm sequentially. Starting from $A^{(0)} = \emptyset$, at step $j \in [k]$, it considers each remaining base arm $i \notin A^{(j-1)}$ and evaluates the reward obtained by adding i to the current partial solution. It then selects the arm with the largest value and continues to the next step. Thus, without elimination, one execution of the greedy oracle requires at most km reward evaluations.

We adapt this procedure to incorporate the active-family constraint in (1). Suppose that the greedy procedure has constructed a partial solution A with $|A| = j - 1$ and considers adding a candidate arm $i \notin A$. To determine whether i can still lead to an active super arm, we complete $A \cup \{i\}$ using the same greedy procedure, but with the optimistic parameter vector $\overline{\boldsymbol{\mu}}_\ell$. More precisely, starting from $B^{(j)} = A \cup \{i\}$, for $q = j + 1, \ldots, k$, choose

$$i_q \in \arg \max_{h \notin B^{(q-1)}} r_{\overline{\boldsymbol{\mu}}_\ell}\big(B^{(q-1)} \cup \{h\}\big), \tag{7}$$

and set $B^{(q)} = B^{(q-1)} \cup \{i_q\}$. Let

$$\widetilde{U}_\ell(A, i) \coloneqq r_{\overline{\boldsymbol{\mu}}_\ell}\big(B^{(k)}\big) \tag{8}$$

denote the optimistic value of this greedy completion.

The role of $\mathcal{C}_\ell$ is now explicit. If

$$\widetilde{U}_\ell(A, i) \geq \Gamma_\ell, \tag{9}$$

then the greedy completion itself provides a super arm $B^{(k)} \in \mathcal{C}_\ell$ containing the current prefix $A \cup \{i\}$. Hence the candidate i can be retained in the greedy construction. Candidates whose greedy optimistic completions fall below the elimination threshold are discarded. Among the retained candidates, we proceed exactly as in the original greedy oracle and select the arm with the largest greedy value.

This implementation avoids enumerating $\mathcal{C}_\ell$. To see its computational cost, consider one greedy step. There are at most m candidate arms. For each candidate, its optimistic completion contains at most k greedy steps, and each such step scans at most m remaining arms. Hence one completion requires $O(km)$ reward evaluations, and testing all candidates at one greedy step requires $O(km^2)$ reward evaluations. Since the construction contains at most k steps, the entire procedure uses at most

$$O(k^2 m^2) \tag{10}$$

reward evaluations per batch. Therefore, as long as one evaluation of $r_{\boldsymbol{\theta}}(S)$ can be performed in polynomial time, the elimination step can also be implemented in polynomial time. Importantly, the algorithm never constructs or stores the potentially exponentially large family $\mathcal{C}_\ell$; it only performs polynomially many greedy reward evaluations.