Avoiding exp(k^*) Scaling for Thompson Sampling in Combinatorial Semi-Bandits: From Multiple Seeds to a Single Seed

Tianyuan Jin[*] TIANYUAN1044@GMAIL.COM
Thrust of Data Science and Analytics,
The Hong Kong University of Science and Technology (Guangzhou)

Heyang Zhao[*] HYZHAO@CS.UCLA.EDU
Department of Computer Science,
University of California, Los Angeles, CA 90095, USA

Vincent Y. F. Tan VTAN@NUS.EDU.SG
Department of Mathematics,
Department of Electrical and Computer Engineering,
National University of Singapore, Singapore 119076

Quanquan Gu QGU@CS.UCLA.EDU
Department of Computer Science,
University of California, Los Angeles, CA 90095, USA

Editors: Steve Hanneke and Tor Lattimore

Abstract

The Combinatorial Multi-Armed Bandit (CMAB) framework extends classical multi-armed bandit theory to complex decision-making settings where agents select super arms to maximize a collective reward. While Thompson Sampling (TS) is widely favored for its robust empirical performance in these settings, its theoretical guarantees have historically suffered from a significant bottleneck: standard Combinatorial Thompson Sampling (CTS) incurs a regret bound with an exponential dependence on the optimal super-arm size k^*. This exponential term arises because standard independent posterior sampling fails to coordinate optimism across the base arms of the optimal super arm, causing the probability of exploration to vanish as k^* increases. Although recent advances have achieved polynomial regret for *linear* rewards, designing an efficient TS algorithm for general, non-linear CMABs remains an open challenge.

In this paper, we resolve this open question by proposing *Combinatorial Thompson Sampling with a Single Seed* ($\texttt{CTS}^3$). Unlike standard approaches that sample base arms independently, $\texttt{CTS}^3$ employs a comonotonic coupling strategy: it generates parameters for all base arms using a single shared random seed via the inverse CDF transform. This mechanism synchronizes sampling fluctuations across arms, ensuring concerted optimism and preventing the exploration probability from decaying exponentially. We prove that $\texttt{CTS}^3$ achieves a regret bound of $O\big(\frac{mkk^*B^2}{\Delta_{\min}}\mathrm{poly}(\log(T, m, \Delta_{\max}/\Delta_{\min}))\big)$ for general reward functions satisfying monotonicity and bounded smoothness, where m is the number of total base arms and k is the largest super arm size. To the best of our knowledge, this is the first polynomial regret bound for Thompson Sampling in general CMAB settings. Empirical evaluations confirm that $\texttt{CTS}^3$ significantly outperforms standard independent TS, particularly in regimes with large super arms.

Keywords: Combinatorial Multi-Armed Bandits, Thompson Sampling, Regret Analysis

$*$. Equal contribution.

1. Introduction

The multi-armed bandit (MAB) (Thompson, 1933) is a paradigmatic model in sequential decision-making, which models the interaction between an agent and an unknown environment over T rounds with arm set $[m] = \{1, 2, \ldots, m\}$. At each round $t \in [T]$, the agent metaphorically pulls an arm based on observations from the first $t-1$ rounds, and then receives a reward drawn from an unknown distribution associated with the chosen arm. We assume each random reward is independent of past selections and outcomes. The objective is to maximize the cumulative reward, or equivalently, to minimize regret, defined as the expected difference in performance between an optimal strategy and the chosen policy.

The combinatorial multi-armed bandit (CMAB) problem (Chen et al., 2013) extends this setting and has received substantial attention due to its applications in computational sociology (Kempe et al., 2003), online advertising (Wang et al., 2017), and resource allocation (Kasy and Teytelboym, 2023). In CMABs, there are m *base arms*, and a *super arm* consisting of a non-empty subset of these base arms is selected at each round. In this paper, we focus on the semi-bandit feedback model, where the agent observes the individual outcome of *every* base arm contained within the selected super arm at the end of each round. We assume that the outcome of base arm i is independently sampled from a Gaussian distribution with mean μ_i and unit variance and we denote the vector of unknown means as $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$.

Among CMAB algorithms, Thompson Sampling (TS) is widely favored for its robust empirical performance. The standard Combinatorial Thompson Sampling (`CTS`) operates by drawing posterior samples $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ to estimate the expected outcomes of the base arms. Given a known reward function $r(S, \boldsymbol{\theta})$, the algorithm selects the super arm $S \subseteq [m]$ that maximizes the reward based on the sampled parameters. Wang and Chen (2018) established that `CTS` achieves a regret bound of $O\big(\frac{m}{\Delta_{\min}} \log T\big) + \tilde{O}\big(\Delta_{\max} \cdot \mathcal{L}^{k^*}\big)$, where $\mathcal{L} = \frac{(k^*)^4}{\Delta_{\min}^2}$, k^* denotes the size of the optimal super arm, and $\Delta_{\min}$ represents the minimum reward gap. While `CTS` yields a matching logarithmic dependence on T and linear dependence on m as its counterparts based on Upper Confidence Bound (UCB) methods (Chen et al., 2016), the exponential term $\mathcal{L}^{k^*}$ in its regret bound is a significant drawback, especially when k^* is large and $\Delta_{\min}$ is small. Recently, Zhang and Combes (2024) proposed a variance-inflated `CTS` variant which effectively reduces the exponential term to a polynomial one. However, their analysis is strictly limited to the linear reward setting, where the reward is the summation of base arm outcomes. Thus, it remains an open problem:

Is it possible to design a Thompson Sampling algorithm with a regret bound that is polynomial in $k^*, 1/\Delta_{\min}$ for general CMABs described in Wang and Chen (2018)?

1.1. Why `CTS` Fails to Remove $\mathcal{L}^{k^*}$

A primary challenge in analyzing `CTS` is the *dependency* introduced by overlapping super arms. Although the outcomes of base arms are independent, the estimated rewards of super arms become statistically coupled because different super arms often share observations from the same set of base arms. To illustrate this, let $S^* = \arg\max_{S \subseteq [m]} r(S, \boldsymbol{\mu})$ denote the optimal super arm, and suppose several suboptimal super arms overlap with S^*. At any given round t, some base arms in S^* may be underestimated while others are overestimated.[1] Under these conditions, the algorithm is biased

1. For simplicity, we assume a super arm's estimated reward decreases if a constituent base arm is underestimated and increases if it is overestimated.

toward selecting super arms containing the overestimated base arms, leaving the underestimated base arms in S^* to be sampled far less frequently. Consequently, certain base arms in S^* can remain severely underestimated for long periods, while others become slightly overestimated due to more frequent sampling.

This dependency is intimately related to the regret term $\mathcal{L}^{k^*}$ that arises in Wang and Chen (2018). Broadly speaking, $\mathcal{L}^{k^*}$ represents the "price" paid to establish an *optimism* property for CTS—specifically, ensuring that the reward of S^* under the posterior sample is within $\Theta(\Delta_{\min})$ of its true mean reward. Achieving this property typically requires the posterior samples θ_i to be concentrated near μ_i for *every* base arm $i \in S^*$, a condition that is significantly harder to satisfy when arms are sampled unevenly due to the aforementioned dependency.

Specifically, Wang and Chen (2018) demonstrated that the *minimum* probability that a base arm $i \in S^*$ is "good"—meaning its posterior sample θ_i falls within an $\Theta(\Delta_{\min}/k^*)$ neighborhood of μ_i—is only on the order of $\Omega(1/\mathcal{L})$. Furthermore, there may be $\Theta(k^*)$ base arms in S^* whose empirical means are simultaneously underestimated; CTS may only select S^* when all samples from these $\Theta(k^*)$ base arms are simultaneously near their true means. Because of the dependency issue, it is possible that at time t, each base arm independently possesses only this minimum probability of being "good." If one requires these events to hold for all $i \in S^*$ simultaneously, a crude product bound suggests that CTS might require on the order of $\mathcal{L}^{k^*}$ pulls of suboptimal super arms before S^* is selected. This logic leads directly to the exponential term $\mathcal{L}^{k^*}$ in their regret bound. Indeed, Wang and Chen (2018) proved the existence of a bandit instance for which CTS incurs regret on the order of $\Omega(2^{k^*})$.

1.2. Our Contributions

In this paper, we resolve the open question by proposing a novel Thompson Sampling algorithm, ***C**ombinatorial **T**hompson **S**ampling with a **S**ingle **S**eed* (CTS^3). Unlike standard CTS, which samples the parameter of each base arm independently from its posterior distribution, CTS^3 instead generates the parameters for all base arms using a single shared random seed. Concretely, at each round, CTS^3 draws one random variable $U \sim \mathrm{Unif}[0,1]$ and then sets the sample for each base arm i to be $F_i^{-1}(U)$, where $F_i(\cdot)$ denotes the CDF of the posterior distribution of base arm i. As a result, this coupling strategy synchronizes sampling fluctuations across arms, thereby inducing concerted optimism; consequently, it prevents the exploration probability from decaying exponentially with the super-arm size.

Building on this construction, we prove that CTS^3 achieves an instance-dependent regret bound of $O\left(\frac{mkk^*B^2}{\Delta_{\min}} \cdot \mathrm{poly}(\log(T, m, \Delta_{\max}/\Delta_{\min}))\right)$. To the best of our knowledge, this is the first polynomial regret bound for Thompson Sampling in CMABs with general reward functions. Here, k denotes the largest super-arm size; see Table 1 for comparisons with TS algorithms. Moreover, we show that the worst-case regret of CTS^3 is $O\left(B\sqrt{mkk^*T} \cdot \mathrm{poly}\big(\log(T, m, \Delta_{\max})\big)\right)$, which matches the worst-case lower bound $\Omega(\sqrt{mkT})$ (Kveton et al., 2015b) up to a $\sqrt{k^*}$ factor (and polylogarithmic terms).

Our analysis introduces several novel techniques (Please see Section 4.1 and our proofs for details) to address the challenges posed by the single-seed coupling, which induces complex dependencies among the sampled parameters of different base arms. We conduct experiments [2] on synthetic datasets,

2. Due to space limit, we defer the experimental results to Appendix B.

Table 1: Comparison with Thompson Sampling Algorithms. The "Ind-Out" column indicates whether the analysis assumes *Independent Outcomes*: conditional on the past, the outcomes of different base arms in the *same round* are independent. "Monotone" refers to Assumption 2.5. For Zhang and Combes (2024), the dependence on $1/\Delta_{\min}$ in $\mathrm{poly}(\cdot)$ is $(1/\Delta_{\min}^2)^{1+\gamma}$ for some $\gamma > 0$.

ALGORITHM	MONOTONE	IND-OUT	REGRET	ADDITIONAL ASSUMPTIONS
WANG AND CHEN (2018)	✗	✓	$O\left(\frac{m \log T}{\Delta_{\min}}\right)+\tilde{O}\left(\mathcal{L}^{k^*}\right)$	2.1, 2.3, AND 2.8
ZHANG AND COMBES (2024)	✓	✓	$\tilde{O}\left(\frac{m^2 k}{\Delta_{\min}}+\mathrm{poly}\left(m, k, \frac{1}{\Delta_{\min}}\right)\right)$	LINEAR REWARDS AND 2.8
CTS^3 (*This work*)	✓	✗	$O\left(\frac{mkk^* B^2 \mathrm{poly}(\log(T, m, \Delta_{\max}/\Delta_{\min}))}{\Delta_{\min}}\right)$	2.1, 2.3, AND 2.8

demonstrating that CTS^3 significantly outperforms standard CTS and other baselines, particularly in scenarios with large super arms.

Notation We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. We use $\mathcal{I}$ for the super arm set, k for the largest super arm size, i.e., $k = \max_{S \in \mathcal{I}} |S|$, S^* for the optimal super arm, and k^* for the size of S^*. We denote by $[n]$ the set $\{1, \ldots, n\}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n = 1, 2, \ldots$, we write $a_n = O(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \leq C b_n$ holds for all $n \geq 1$ and write $a_n = \Omega(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \geq C b_n$ holds for all $n \geq 1$. We use $\tilde{O}(\cdot)$ to further hide the polylogarithmic factors. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function. For any scalar x, write $x^+ = \max\{x, 0\}$. We denote $\Delta_{\min} = \min_{S: \Delta_S > 0} \Delta_S$ and $\Delta_{\max} = \max_{S \in \mathcal{I}} \Delta_S$.

2. Preliminaries

In this section, we formally define the combinatorial semi-bandit problem and introduce some notations used in the paper.

2.1. Combinatorial Semi-Bandits

In Combinatorial Multi-Armed bandits (CMAB), we have a set of base arms $[m] = \{1, 2, \cdots, m\}$ and the super arm set $\mathcal{I}$. At each round t, each base arm i has marginal outcome $Z_i(t) \sim \mathcal{N}(\mu_i, 1)$, and the marginal outcome process of each arm is independent over time conditional on the past. Let $\boldsymbol{Z}(t) = \{Z_1(t), Z_2(t), \cdots, Z_m(t)\}$ be the random outcomes of all base arms. Importantly, we *do not* require the coordinates of $\boldsymbol{Z}(t)$ to be independent within the same round: $\boldsymbol{Z}(t)$ may have arbitrary contemporaneous correlation across arms. Based on the past observations, the agent pulls a super arm $S(t) \in \mathcal{I}$ and receives feedback at the end of round t. The expected reward of $S(t)$ is denoted by $R(S(t))$ and the feedback is represented by $Q(t) = Q(S(t), \boldsymbol{Z}(t))$. In this paper, we consider the semi-bandit feedback, i.e., $Q(t) = \{(i, Z_i(t)) : i \in S(t)\}$.

For the reward function, we make some commonly used assumptions here. We let $\boldsymbol{\mu}_S$ be the projection of $\boldsymbol{\mu}$ on S, i.e., $\boldsymbol{\mu}_S$ is $\boldsymbol{\mu}$ restricted to S and zero elsewhere.

Assumption 2.1 *The expected reward of S only depends on the mean outcomes. That is, there exists a function r such that*

$$\mathbb{E}[R(S(t))] = r(S, \boldsymbol{\mu}) = r(S, \boldsymbol{\mu}_S).$$

Remark 2.2 *Assumption 2.1 is widely used in CMAB literature (Chen et al., 2013, 2016; Wang and Chen, 2018). It indicates that the expected reward of a super arm S is determined solely by the mean outcomes of the base arms in S, and is independent of the specific distributions of the outcomes.*

This assumption holds for many common reward functions, such as linear rewards (Gai et al., 2012; Kveton et al., 2015b) and probabilistic maximum coverage rewards (Chen et al., 2016).

Example 1 (Linear reward) *For linear reward functions, due to the linearity of expectation, the expected reward of a sum is always the sum of the expected rewards, regardless of whether the arms are correlated or independent. The reward is simply the sum of the outcomes of the selected arms. Since $\mathbb{E}[R(S, \mathbf{Z})] = \sum_{i \in S} \mathbb{E}[Z_i] = \sum_{i \in S} \mu_i$, the expected reward depends only on the mean outcomes of the selected arms.*

Example 2 (Probabilistic maximum coverage) *Probabilistic maximum coverage (Chen et al., 2016) is a classic non-linear example used in viral marketing under the standard independent Bernoulli-outcome CMAB model. This example illustrates the mean-dependent reward structure, although our main theorem below is stated for Gaussian semi-bandit observations. It models a situation where you pick a seed set of users, and you want to maximize the number of people they influence. At each round, the learner selects a set S of base arms (seed nodes). Each node $i \in S$ has a probability μ_i (its mean) of "activating" or influencing a target. The reward is 1 if the target is influenced by at least one node in S, and 0 otherwise. Under independent Bernoulli activations,*

$$\mathbb{E}[R(S)] = \mathbb{E}\left[1 - \prod_{i \in S}(1 - Z_i)\right] = 1 - \mathbb{E}\left[\prod_{i \in S}(1 - Z_i)\right] = 1 - \prod_{i \in S}(1 - \mathbb{E}[Z_i]) = 1 - \prod_{i \in S}(1 - \mu_i),$$

where Z_i is a binary outcome: 1 if node i activates, 0 otherwise.

We next impose an ℓ_1-Lipschitz regularity condition on the reward mapping r to handle potentially non-linear reward functions. For any vector $x = (x_1, \ldots, x_m)$, we write $\|x\|_1 = \sum_{i \in [m]} |x_i|$ for its ℓ_1 norm.

Assumption 2.3 (Bounded Smoothness) *There is some constant $B > 0$ such that for all pairs of mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$,*

$$|r(S, \boldsymbol{\mu}) - r(S, \boldsymbol{\mu}')| \leq B \|\boldsymbol{\mu}_S - \boldsymbol{\mu}'_S\|_1.$$

Remark 2.4 *Assumption 2.3 is a standard regularity condition in the CMAB literature, often referred to as **1-Norm Bounded Smoothness** (Chen et al., 2013). It generalizes the specific structure of strictly linear rewards while maintaining control over the reward function's sensitivity.*

We further make the following assumption which is also popular in the combinatorial semi-bandits (Chen et al., 2013, 2016; Kveton et al., 2015b,a; Chen et al., 2016).

Assumption 2.5 (Monotone) *The expected reward of any super arm $S \in \mathcal{I}$ is a non-decreasing function with respect to the mean vector. That is, if for all $i \in [m]$, it holds that $\mu_i \leq \mu_i'$, then*

$$r(S, \boldsymbol{\mu}) \leq r(S, \boldsymbol{\mu}') \quad \textit{for all } S \in \mathcal{I}.$$

Remark 2.6 *The monotonicity assumption reflects the intuitive property that improving the quality of individual base arms (i.e., increasing their expected means μ_i) should not decrease the total expected reward of any super arm. This condition holds for virtually all standard combinatorial applications. In network reliability problems, improving the success probability of an individual edge strictly increases the global connectivity probability (Gai et al., 2012). In crowdsourcing, the collective accuracy of a worker pool (e.g., under majority voting) is monotonic with respect to individual worker accuracies (Karger et al., 2014). Furthermore, in assortment optimization under the Multinomial Logit model, increasing the preference weight of a product typically yields higher expected total revenue (Agrawal et al., 2019).*

Remark 2.7 *The monotonicity assumption is a cornerstone of the Generalized CMAB framework, enabling the analysis of complex, non-linear reward structures. It was originally formalized by Chen et al. (2013) to bridge the gap between linear bandits and influence maximization problems. Since then, it has been adopted as a standard requirement in all general CMAB analyses, including Merlis and Mannor (2019), Chen et al. (2013) and Chen et al. (2016). It is worth noting that Wang and Chen (2018) supported non-monotone reward functions in their analysis of `CTS`; however, this generality came at the cost of introducing an exponential term in the regret bound, as discussed in our Section 1 and their proposed lower bound in their Theorem 3.*

Moreover, we make the following assumption that is crucial for Thompson Sampling (Wang and Chen, 2018).

Assumption 2.8 (Greedy Oracle, Wang and Chen 2018) *The algorithm can exactly access the oracle* ORACLE($\boldsymbol{\theta}$) *that takes $\boldsymbol{\theta}$ as input, and outputs a super arm* $\operatorname{argmax}_{S \in \mathcal{I}} r(S, \boldsymbol{\theta})$*, where $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_m\}$.*

Remark 2.9 (Necessity of Exact Oracles) *Unlike UCB-based approaches which are robust to approximation errors (Chen et al., 2016), standard Thompson Sampling can suffer **linear regret** $\Omega(T)$ when paired with a deterministic approximation oracle. As illustrated by the counter-example in Wang and Chen (2018), an approximation oracle may satisfy its multiplicative guarantee (e.g., returning a solution within factor λ of the optimal) by consistently selecting a suboptimal super-arm that is merely "good enough," even when the sampled parameters for the optimal super-arm are optimistic. Consequently, the optimal super-arm is* never selected *and* never observed*, causing its posterior distribution to remain frozen at the prior. This prevents the algorithm from ever learning the true quality of the optimal arm, leading to a permanent failure in exploration. Therefore, the assumption of an **Exact Oracle** is not merely for convenience but is a structural requirement for the consistency of standard Combinatorial Thompson Sampling.*

Objective. Our goal is to minimize the regret, which is defined below:

$$R_{\mathcal{I}}(T) = \sum_{t=1}^{T} \big(r(S^*, \boldsymbol{\mu}) - r(S(t), \boldsymbol{\mu}) \big),$$

where $S^* = \arg\max_{S \in \mathcal{I}} r(S, \boldsymbol{\mu})$.

2.2. Discussion on Possible Solutions to Remove the $\mathcal{L}^{k^*}$ Term

In this subsection, we discuss potential approaches in the literature that may remove the $\mathcal{L}^{k^*}$ term. Due to space constraints, we defer further related work to Appendix A.

Inflating the variance. As discussed in Section 1, the term $\mathcal{L}^{k^*}$ in the regret analysis of CTS stems from the *underestimation* of the optimal super arm S^*.

A common remedy in Thompson Sampling is to *inflate the sampling variance*, which is powerful in solving the underestimation for MAB (Jin et al., 2021) and for linear bandits (Agrawal and Goyal, 2012a). Intuitively, when a base arm is underestimated, drawing a more dispersed posterior sample increases the probability that the sample becomes optimistic (i.e., not too far below the true mean), hence the algorithm may still pull that base arm in CTS.

However, variance inflation does not fundamentally solve the issue: to select S^*, the oracle typically needs many coordinates in S^* to appear simultaneously "good" (Recall good for base arm i means that the posterior sample θ_i falls within an $O(\Delta_{\min}/k^*)$ neighborhood of μ_i)[3]. Thus, even if inflation increases the minimum probability of being optimistic from $\Theta(\Delta_{\min}^2)$ to some universal constant $c > 0$, the probability that k^* base arms are optimistic *at the same time* can still scale as c^{k^*}, i.e., exponentially small in k^*. Moreover, for symmetric posteriors such as Gaussian, this constant cannot be larger than $1/2$: for any Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{P}(Z \geq \mu) = 1/2$, regardless of σ^2. Therefore, variance inflation can at best improve an exponential factor from $\mathcal{L}^{k^*}$ to 2^{k^*}, but it still leaves an exponential dependence on k^*.

For *special structures*, variance inflation can remove the exponential term. For example, in *linear* combinatorial bandits where the reward is $r(S, \boldsymbol{\mu}) = \sum_{i \in S} \mu_i$, one can inflate the sampling variance and obtain a regret bound with only polynomial dependence on k^*. Intuitively, inflating the sampling variance boosts the chance that some $\theta_i(t)$ for $i \in S^*$ is significantly above μ_i. Due to the linearity of the reward function, this single optimistic sample can compensate for others, meaning the estimates for the remaining base-arm samples only need to be close to the ground-truth (rather than optimistic). Consequently, the resulting regret depends only polynomially on k^*, instead of exponentially. Specifically, for linear rewards, BG-CTS (Zhang and Combes, 2024) samples $\theta_i(t) \sim \mathcal{N}\big(\hat{\mu}(t), 2g(t)/N_i(t)\big)$, so the sampling variance is inflated by a factor $2g(t)$, where $g(t) = O(k \log\log t)$. With this choice, the expected regret bound takes the form

$$O\big(\frac{m \ln k}{\Delta_{\min}} \ln T\big) + O\big(\frac{m^2 k \ln k}{\Delta_{\min}} \ln\ln T\big) + \mathrm{poly}\big(m, k, \frac{1}{\Delta_{\min}}\big).$$

In particular, the dependence on $1/\Delta_{\min}$ in $\mathrm{poly}(\cdot)$ is $(1/\Delta_{\min}^2)^{1+\gamma}$ for some $\gamma > 0$.

Correlated sampling. Correlated sampling (Agrawal et al., 2017) is also known for solving the *underestimation* of optimal arm in MNL-Bandit. The idea is simple: at each round t, draw k samples from $\mathcal{N}(0, 1)$, denoted by $\{\theta_i^{(j)}(t)\}_{j=1}^k$, and then form an "optimistic" sample by

$$\theta_i(t) = \max_{j \in [k]} \hat{\mu}_i(t-1) + \frac{\eta \theta_i^{(j)}(t)}{\sqrt{N_i(t-1)+1}}.$$

This coordinate-wise maximization and the variance inflation parameter η increases the probability that the sample of an underestimated optimal arm to be large and hence can be selected more often.

3. For ease of presentation, we assume $B = 1$ in the main paper. Then, if the posterior samples satisfy $\theta_i \in \mu_i \pm O(\Delta_{\min}/k)$ for all $i \in S^*$, it follows that $r(S^*, \boldsymbol{\theta}) \geq r(S^*, \boldsymbol{\mu}) - O(\Delta_{\min})$. .

Zhong et al. (2021) study *cascading bandits* and propose a variant of TS algorithm based on a *shared seed* construction. Their shared-seed idea is used to *save a* $\sqrt{k}$ *factor* in the regret bound. However, the designed algorithm and the regret analysis crucially depends on the cascading structure and the Gaussian construction (with additional truncation to match Bernoulli clicks), and does *not* extend to CTS with independently sampled base arms (they explicitly note the key step fails for CTS in the discussion after the proof of their Lemma 4.4).

It is unclear whether the above tricks can be applied to *combinatorial semi-bandits*. Establishing theoretical guarantees for such a correlated-sampling variant of CTS remains open, and is explicitly raised as an open problem in Wang and Chen (2018).

More importantly, we also note that correlated sampling schemes are often paired with aggressive variance inflation to further promote optimism (Agrawal et al., 2017; Zhang and Combes, 2024), which may worsen the empirical performance.

Our aim. In this paper, we aim to design a CTS algorithm that *preserves the exact posterior* marginal of each base arm—without inflating variances or otherwise modifying any arm's marginal sampling distribution. Moreover, we seek to eliminate the $\mathcal{L}^{k^*}$ factor that appears in existing regret bounds of CTS for general CMABs. Finally, we aim to preserve the simplicity of CTS while maintaining—or even improving—its empirical performance.

3. Single-Seed Thompson Sampling

In this section, we present ***C**ombinatorial **T**hompson **S**ampling with a **S**ingle **S**eed* (CTS^3), whose pseudo-code is given in Algorithm 1. CTS^3 modifies standard combinatorial Thompson sampling by *coupling* all coordinate-wise posterior samples through one common random seed. After initialization, every base arm has been observed at least once; hence $N_i(t-1) \geq 1$ for all Thompson-sampling rounds $t \geq t_0 + 1$. Specifically, for each base arm $i \in [m]$, the algorithm maintains a Gaussian posterior $\mathcal{N}\big(\hat{\mu}_i(t-1), 1/N_i(t-1)\big)$, where $\hat{\mu}_i(t-1)$ is the empirical mean and $N_i(t-1)$ is the number of observations of arm i up to round $t-1$. At each round $t \geq t_0 + 1$, CTS^3 draws a single seed $X_t \sim \mathcal{N}(0,1)$ and forms the sample

$$\theta_i(t) = \hat{\mu}_i(t-1) + \frac{X_t}{\sqrt{N_i(t-1)}}, \qquad \forall i \in [m].$$

Thus $\theta_i(t)$ has marginal distribution $\mathcal{N}(\hat{\mu}_i(t-1), 1/N_i(t-1))$. Throughout the analysis, this Gaussian posterior is the posterior under the improper flat prior $p(\mu_i) \propto 1$; after $s \geq 1$ observations from arm i, the posterior is $\mathcal{N}(\hat{\mu}_{i,s}, 1/s)$. The initialization phase ensures that this posterior is proper before Thompson sampling starts. Let $\boldsymbol{\theta}(t) = (\theta_1(t), \ldots, \theta_m(t))$. Given $\boldsymbol{\theta}(t)$, the algorithm calls the oracle to select the super arm $S(t) = \textsc{Oracle}(\theta(t))$, plays $S(t)$, and observes semi-bandit outcomes $\{Z_i(t) : i \in S(t)\}$. It then updates $(\hat{\mu}_i(t), N_i(t))$ for each $i \in S(t)$ accordingly.

Generalization to other reward distributions. While we present CTS^3 for Gaussian outcomes, the single seed idea extends to other reward models. Specifically, suppose each base arm has a parametric reward distribution indexed by its mean μ_i and we can maintain a posterior over μ_i given the observations. Let $F_{i,t}$ be the CDF of this posterior at round t. Then we can draw a single seed $U_t \sim \mathrm{Uni}(0,1)$ and set $\theta_i(t) = F_{i,t}^{-1}(U_t)$ for all $i \in [m]$. For exponential-family rewards with conjugate priors (e.g., Bernoulli–Beta), $F_{i,t}^{-1}$ is available in standard libraries; otherwise one may use numerical inversion or a Laplace/Gaussian approximation.

Algorithm 1 **C**ombinatorial **T**hompson **S**ampling with A **S**ingle **S**eed (CTS^3)

Input: Super Arm Set $\mathcal{I}$; ORACLE() as defined in Assumption 2.8.
Initialization: Set $N_i(0) \leftarrow 0$ and $\hat{\mu}_i(0) \leftarrow 0$ for all $i \in [m]$. Play a sequence of super arms $S(1), S(2), \ldots, S(t_0) \in \mathcal{I}$ such that every base arm is observed at least once, i.e., $N_i(t_0) \geq 1, \forall i \in [m]$, and choose such a sequence with $t_0 \leq m$. For each $t = 1, \ldots, t_0$, after playing $S(t)$ and observing $\{Z_i(t) : i \in S(t)\}$, update $N_i(t)$ and $\hat{\mu}_i(t)$ accordingly.
for $t = t_0 + 1, t_0 + 2, \ldots$ **do**
  Sample a common seed $X_t \sim \mathcal{N}(0, 1)$;
  for $i = 1, 2, \ldots, m$ **do**
    $\theta_i(t) \leftarrow \hat{\mu}_i(t-1) + X_t / \sqrt{N_i(t-1)}$;
  end for
  Let $\boldsymbol{\theta}(t) \leftarrow (\theta_1(t), \ldots, \theta_m(t))$;
  $S(t) \leftarrow$ ORACLE$(\boldsymbol{\theta}(t))$;
  Play $S(t)$ and observe outcomes $\{Z_i(t) : i \in S(t)\}$;
  for each $i \in S(t)$ **do**
    $N_i(t) \leftarrow N_i(t-1) + 1$;
    $\hat{\mu}_i(t) \leftarrow \frac{N_i(t-1)\hat{\mu}_i(t-1) + Z_i(t)}{N_i(t)}$;
  end for
  for each $i \notin S(t)$ **do**
    $N_i(t) \leftarrow N_i(t-1)$;
    $\hat{\mu}_i(t) \leftarrow \hat{\mu}_i(t-1)$;
  end for
end for

4. Regret Analysis of CTS^3

In this section, we first present the main technical challenges in the regret anlysis of CTS^3, then we provide main results for CTS^3.

4.1. Technical Challenges in Regret Analysis and the Proof Outline

This section explains (i) why the standard regret analysis of CTS in Wang and Chen (2018) does not extend to CTS^3, (ii) what new technical challenges are created by the single-seed coupling, and (iii) the main proof ideas used to obtain our regret bound. For ease of presentation, we assume $B = 1$.

Before we start, we define some notations for ease of presentation. For each integer $r \in \mathbb{Z}$, define the bucket $M_r = \left\{S \in \mathcal{I} : \ \Delta_S \in (2^{-r}, 2^{-r+1}]\right\}$, and let $M_\infty = \{S \in \mathcal{I} : \Delta_S = 0\}$ be the set of optimal super arms. For a nonempty bucket M_r, define the post-initialization bucket regret

$$\mathbb{E}[R_{M_r}(T)] = \mathbb{E}\Big[\sum_{t=t_0+1}^{T} \Delta_{S(t)} \mathbb{1}\{S(t) \in M_r\}\Big].$$

We focus on bounding $R_{M_r}(T)$. Let $F_{i,t}$ denote the CDF of $\mathcal{N}(\hat{\mu}_i(t-1), 1/N_i(t-1))$, which is the posterior distribution of arm i at Thompson-sampling round t. Set $\rho_T = (Tmk)^{-8}$ and define the lower/upper ρ_T-quantiles $\theta_{\min,i}(t) = F_{i,t}^{-1}(\rho_T)$ and $\theta_{\max,i}(t) = F_{i,t}^{-1}(1-\rho_T)$. Let $\delta_r = \frac{1}{8}\min_{S \in M_r} \Delta_S$ when $B = 1$. Define $W_r(t)$ to be the set of super arms that are not overestimated at

time t:

$$W_r(t) = \Big\{ S \in M_r : \sum_{i \in S} \big(\theta_{\max,i}(t) - \mu_i\big)^+ \le \delta_r \Big\}. \tag{4.1}$$

Then, up to the initialization regret, the regret caused by pulling arms in M_r can be decomposed into two terms:

$$\mathbb{E}[R_{M_r}(T)] \le 16\delta_r \underbrace{\sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathbb{1}\{S(t) \in M_r, S(t) \notin W_r(t)\}\big]}_{\heartsuit} + 16\delta_r \underbrace{\sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathbb{1}\{S(t) \in M_r, S(t) \in W_r(t)\}\big]}_{\spadesuit}.$$

Intuitively, Term $\heartsuit$ captures the regret caused by selecting a bucket arm because some super arm in M_r is overestimated, while Term $\spadesuit$ is the hard part: $S(t) \in W_r(t)$ means we cannot blame large overestimation, so the mistake comes from underestimation on S^*.

Bounding Term $\heartsuit$. If $S(t) \in M_r$ but $S(t) \notin W_r(t)$, then by definition $\sum_{i \in S(t)} (\theta_{\max,i}(t) - \mu_i)^+ > \delta_r$. On the standard high-probability events, once a base arm i has been pulled at least $\tau_r = O(k^2 \log(mT)/\delta_r^2)$ times, its posterior quantiles cannot deviate above μ_i by more than about δ_r/k. Thus a violation of $W_r(t)$ forces a positive dyadic score, which can be charged to under-sampled base arms. The refined dyadic charging argument gives a contribution of order $m\tau_r \log(2k)/k$ pulls; see Lemma C.5.

Bounding Term $\spadesuit$ (main hardness). We now explain why the usual frequentist analysis template for CTS does not extend to CTS^3, and then give our proof outline in bounding term $\spadesuit$.

(i) How the standard CTS analysis works. In standard CTS, the samples $\{\theta_i(t)\}_{i=1}^m$ are drawn *independently* across base arms. A common method in Thompson Sampling analyses (Agrawal and Goyal, 2017; Jin et al., 2024, 2023; Wang and Chen, 2018) is to introduce a set $\mathcal{S}^o(t)$ of "exploration-enforcing" super arms: those that contain at least one base arm $i \in S^*$ that is *not sufficiently pulled* (so its empirical mean and the posterior sample has not yet concentrated). Then one relates the probability of choosing from $\mathcal{S}^o(t)$ to an "optimism" event via

$$\mathbb{P}(S(t) \in \mathcal{S}^o(t)) \;\le\; \frac{1}{\mathbb{P}(S^* \text{ is optimistic under } \boldsymbol{\theta}(t))} \mathbb{P}\big(S(t) \in M_r \setminus \mathcal{S}^o(t)\big). \tag{4.2}$$

Intuitively, the inequality above shows that we can control the number of times CTS pulls a suboptimal super arm from the bucket M_r by relating it to the frequency with which S^* is optimistic. Recall that $\mathcal{S}^o(t)$ consists of the super arms that contain at least one *under-explored* base arm in S^* (i.e., some $i \in S^*$ with $N_i(t-1) < \tau_r$). Pulling a super arm in $\mathcal{S}^o(t)$ necessarily increases the pull count of at least one such under-explored base arm. Therefore, this can occur only until every base arm in S^* has been pulled at τ_r times. Consequently, the total number of pulls from M_r can be bounded—up to the multiplicative factor $1/\mathbb{P}\big(S^* \text{ is optimistic}\big)$—by the total number of pulls from $\mathcal{S}^o(t)$, which is at most on the order of $k\tau_r$.

However, as discussed in the introduction, the reciprocal of the optimism probability, namely $1/\mathbb{P}(S^* \text{ is optimistic under } \boldsymbol{\theta})$, can be exponentially large, which leads to an $\mathcal{L}^{k^*}$ factor in the regret bound of CTS. While the inverse optimism probability for CTS^3 is not large, a natural question is whether the above proof framework can still be used to bound $\spadesuit$. The following result answers this question in the negative.

(ii) Why the same template breaks for CTS^3. With independent posterior samples in CTS, optimism of S^* often "pays you back": conditioning on S^* being optimistic typically increases the probability that the selected super arm contains at least one not-yet-well-explored base arm in S^*, so CTS is forced to increase the pull counts of those under-explored arms. This intuition can fail under a single-seed coupling. Indeed, when $\theta_i(t) = F_{i,t}^{-1}(U_t)$ is generated from a common seed $U_t \sim \mathrm{Uni}[0,1]$, making $\theta_i(t)$ large for an underestimated arm $i \in S^*$ typically requires U_t to fall in an upper-tail region. Since each $F_{j,t}^{-1}(\cdot)$ is nondecreasing, the same realization of U_t simultaneously pushes many other coordinates $\theta_j(t)$ upward as well. Consequently, under such an U_t, the rewards of *all* super arms tend to increase, rather than only those containing the underestimated arms in S^*. As a result, for CTS^3, the relation in (4.2) no longer holds.

(iii) Our solution: rescue events, seed slicing, and anti-concentration via peeling. Fix r and define

$$\mathcal{H}_1 = M_\infty \cup \bigcup_{a\in\mathbb{Z}:\, a\geq r+2} M_a, \qquad \mathcal{H}_2 = \bigcup_{a\in\mathbb{Z}:\, a\leq r+1} M_a.$$

Define the lower-quantile event

$$\mathrm{Good}_{\mathcal{H}_1}^{\min}(t) = \left\{\exists S \in \mathcal{H}_1 : \sum_{i\in S} |\theta_{\min,i}(t) - \mu_i| \leq \Theta(\delta_r)\right\}.$$

Key implication. If $\mathrm{Good}_{\mathcal{H}_1}^{\min}(t)$ holds, then there exists some $\bar{S} \in \mathcal{H}_1$ whose samples are not severely underestimated (with high probability $\theta_i(t) \geq \theta_{\min,i}(t)$). Under the restriction $S(t) \in W_r(t)$, such a $\bar{S}$ dominates every $S \in M_r$, and hence the oracle cannot output any $S(t) \in M_r$. Therefore, the contribution of Term ♠ is essentially incurred only on rounds where $\mathrm{Good}_{\mathcal{H}_1}^{\min}(t)$ *fails*.

Motivated by this, define $G_t = \neg\mathrm{Good}_{\mathcal{H}_1}^{\min}(t)$ and the random set of critical rounds $L = \{t \in [T] :\ \mathbb{1}\{G_t\} = 1\}$. Then Term ♠ reduces to controlling $\mathbb{E}[|L|]$. Define the "rescue" event that the base arms in S^* are simultaneously not underestimated in *the actual sample*:

$$\mathcal{E}_{\mathrm{under}}(t) \ = \ \bigcap_{i\in S^*} \{\theta_i(t) \geq \mu_i - \Theta(\delta_r/k^*)\}. \tag{4.3}$$

We further split $L = L_1 \cup L_2 \cup L_3$, where $L_1 = \{t \in L : \mathbb{1}\{\mathcal{E}_{\mathrm{under}}(t), S(t) \in \mathcal{H}_1\} = 1\}$, $L_2 = \{t \in L : \mathbb{1}\{S(t) \in \mathcal{H}_2, \mathcal{E}_{\mathrm{under}}(t)\} = 1\}$, and $L_3 = L \setminus (L_1 \cup L_2)$. The sets L_1 and L_2 are *chargeable*: each round $t \in L_1 \cup L_2$ increases the pull count of at least one under-explored base arm. In particular, (i) if every base arm in $\mathcal{H}_1$ has been pulled at least τ_r times, then $\mathrm{Good}_{\mathcal{H}_1}^{\min}(t)$ holds and hence $t \notin L_1$; and (ii) If every base arm in $\mathcal{H}_2$ has been pulled at least τ_r times, then for any $S \in \mathcal{H}_2$ the reward estimate $r(S, \boldsymbol{\theta}(t))$ concentrates around $r(S, \boldsymbol{\mu})$. Hence, under $\mathcal{E}_{\mathrm{under}}(t)$ we have $r(S^*, \boldsymbol{\theta}(t)) \ \geq \ \max_{S\in\mathcal{H}_2} r(S, \boldsymbol{\theta}(t))$, so the selected super arm cannot lie in $\mathcal{H}_2$, i.e., $S(t) \notin \mathcal{H}_2$. Therefore $t \notin L_2$.

These observations already imply the crude bound $|L_1|, |L_2| \leq m\tau_r$. In our proof, we further refine this charging argument by controlling the pull counts at *dyadic scales* and strengthening the associated concentration events, which yields the sharper bound $|L_1|, |L_2| \leq \frac{m\tau_r \log(2k)}{k}$, see Lemma C.1 for details. This part is *one of our main technical novelties*.

The remaining set L_3 corresponds to rounds where all base arms in $S(t)$ are already well explored. Let $p_t = \mathbb{P}\big(\mathcal{E}_{\mathrm{under}}(t) \mid \mathcal{F}_{t-1},\ G_t\big)$. A key observation is that if $\mathcal{E}_{\mathrm{under}}(t)$ holds and $t \in L$, then $t \in L_1 \cup L_2$. Then, among the critical rounds, only a p_t-fraction are "rescued" and contribute to

$L_1 \cup L_2$. Equivalently, each time we count one critical round, we can charge it to a rescued round with an inverse-probability weight $1/p_t$. This yields the inequality $\mathbb{P}\big(G_t \mid \mathcal{F}_{t-1}\big) \le \frac{1}{p_t}\,\mathbb{P}\big(G_t, \mathcal{E}_{\text{under}}(t) \mid \mathcal{F}_{t-1}\big)$, and summing over t gives $\mathbb{E}[|L|] \le \mathbb{E}\big[\sum_{t\in L_1\cup L_2} 1/p_t\big]$.

Final step (the main technical challenges and novelties): controlling $\mathbb{E}\big[\sum_{t\in L_1\cup L_2} \frac{1}{p_t}\big]$ via time-uniform anti-concentration + peeling. Let $\underline{p} := \min_{t\in L} p_t$. Since $\underline{p} \le p_t$ for all $t \in L$, we have $\sum_{t\in L_1\cup L_2} \frac{1}{p_t} \le \frac{|L_1\cup L_2|}{\underline{p}}$, and by the charging argument $|L_1 \cup L_2| \le m\tau_r \log(2k)/k$, it suffices to control $\mathbb{E}[1/\underline{p}]$.

A naive approach is to plug in a uniform worst case lower bound on $\underline{p}$ (e.g., through a crude bound on $\min_{i\in S^*} \min_{t:N_i(t)<\tau_r} \mathbb{P}(\theta_i(t) \ge \mu_i - \Theta(\delta_r/k) \mid \mathcal{F}_{t-1})$). This typically bounds lower bound $\underline{p}$ by $\delta_r^2/\log T$ and leads to an overly loose δ_r^{-3}-type regret bound.

Our analysis avoids this issue in two steps. (i) We first derive an anti-concentration bound for $\mathbb{P}\big(\theta_i(t) \ge \mu_i - \Theta(\frac{\delta_r}{k}) \,\big|\, \hat{\mu}_i(t-1) = z\big)$, and then apply a *peeling* argument to obtain a sharp, time-uniform concentration bound for $\mathbb{P}\big(\exists s \le \tau_r : \hat{\mu}_{i,s} < z\big)$ (see Lemma C.4), where $\hat{\mu}_{i,s}$ denotes the empirical mean of arm i after s pulls. (ii) We further peel over the value of $\underline{p}$ (equivalently, over the corresponding threshold z) using a geometric partition of $(0, 1]$.

Specifically, to bound $\mathbb{E}[1/\underline{p}]$, we peel over the value of $\underline{p}$ by applying a geometric partition of $(0, 1]$ into a collection of slices. The key input in bound $1/p$ in each slices is the Lemma C.4 (time-uniform concentration on the empirical mean): it implies that $\underline{p}$ cannot remain extremely small for many rounds, because the event $\{\underline{p} \in (\epsilon, 2\epsilon)\}$ would force (via Gaussian posterior tails) the base arm to have a low empirical mean at some time $s \le \tau_r$, which Lemma C.4 rules out uniformly over time. Summing the slice-wise contributions then yields $\mathbb{E}[1/\underline{p}] = \tilde{O}(k^*)$, and hence $\mathbb{E}\Big[\sum_{t\in L_1\cup L_2} \frac{1}{p_t}\Big] \le 2m\log(2k)\cdot\tau_r/k\cdot\tilde{O}(k^*)$, which leads to a polynomial bound on $\mathbb{E}[|L|]$ and thus on Term ♠.

4.2. Main Theorem

Theorem 4.1 *Assume that Assumptions 2.1, 2.3, 2.5, and 2.8 hold. Then the regret of* CTS^3 *satisfies*

$$\mathbb{E}[R_{\mathcal{I}}(T)] \le 2m\Delta_{\max} + O\left(\frac{mkk^*B^2}{\Delta_{\min}}\text{poly}(\log(T, m, \Delta_{\max}/\Delta_{\min}))\right). \tag{4.4}$$

Moreover,

$$\mathbb{E}[R_{\mathcal{I}}(T)] \le 2m\Delta_{\max} + O\big(B\sqrt{mk^*kT}\cdot\text{poly}(\log(T, m, \Delta_{\max}))\big). \tag{4.5}$$

Comparison with prior work. Compared with Wang and Chen (2018), our analysis removes the *exponential* dependence on $1/\Delta_{\min}$ in the instance-dependent regret bound. Moreover, compared with recent results for *linear reward* models (Zhang and Combes, 2024), we sharpen the gap dependence from $(1/\Delta_{\min}^2)^{1+\gamma}$ to an essentially near-linear dependence on $1/\Delta_{\min}$ (up to polylogarithmic factors). As a consequence, our bound yields a worst case regret of order $\sqrt{T}$ (up to polylogarithmic factors) as a function of the horizon T. In contrast, a gap dependence of the form $(1/\Delta_{\min}^2)^{1+\gamma}$ in Zhang and Combes (2024) typically leads to the worst case regret bound $\max_{\Delta_{\min}>0}\Big\{\min\{T\Delta_{\min}, (1/\Delta_{\min}^2)^{1+\gamma}\}\Big\} = \Theta\big(T^{\frac{2+2\gamma}{3+2\gamma}}\big)$, which is at least $T^{2/3}$ as $\gamma > 0$.

Near optimal on worst case regret bound. Kveton et al. (2015b) establish a worst case regret lower bound: for any algorithm, there exists a CMAB instance on which the regret is $\Omega(\sqrt{mkT})$. Compared with this lower bound, our worst case regret upper bound is within a $\sqrt{k^*}$ factor, up to polylogarithmic terms, of the worst case optimal.

Remark 4.2 (Discussion on the additional factor $\sqrt{k^*}$.) *For the worst case regret, our bound is within a $\sqrt{k^*}$ factor of the worst case optimal rate (up to polylogarithmic terms). This $\sqrt{k^*}$ gap mainly comes from applying a union bound over $i \in S^*$ for Lemma C.4 when we control $\underline{p} := \min_{t \in L} p_t$, which introduces an extra $\sqrt{k^*}$ factor. Removing this $\sqrt{k^*}$ for Thompson sampling appears nontrivial, and we leave it as an open problem. We conjecture that inflating the posterior variance by a factor of $\log k$ could remove the $\sqrt{k^*}$ term. However, we do not pursue this direction, as it deviates from our aim of sampling from the true marginal posterior and may degrade empirical performance.*

5. Conclusion and Future Work

We studied combinatorial bandits with semi-bandit feedback. We proposed *Combinatorial Thompson Sampling with a Single Seed* (CTS^3). Like CTS, CTS^3 samples from the posterior of each base arm. The key difference is the sampling *coupling*: CTS draws these posterior samples independently across arms, while CTS^3 ties them together using a single shared random seed. We established a *polynomial* instance-dependent regret bound $\tilde{O}\big(\frac{mkk^*}{\Delta_{\min}}\big)$ for CTS^3, along with a $\tilde{O}(\sqrt{T})$-type worst case bound. Empirically, on the stochastic maximum spanning tree task, CTS^{3}consistently improves over standard CTSand competitive UCB baselines, especially when the base arm size is large. [4]

Future work. Our work opens several directions:

- **Closing the $\sqrt{k^*}$ gap.** Our worst-case bound is within a $\sqrt{k^*}$ factor (up to polylogarithmic terms) of the known worst case lower bound. Whether or not we can remove this $\sqrt{k^*}$ factor in CTS^3 is a natural and technically interesting open problem.
- **Weaker oracles.** Our analysis assumes access to an exact optimization oracle. It would be valuable and interesting to develop and analyze variants that work with approximate oracles.

Acknowledgment

We thank the anonymous reviewers and area chair for their helpful comments. TJ and VYFT are supported by a Singapore Ministry of Education (MOE) AcRF Tier 2 grant under grant number A-8004062-00-00. HZ and QG are supported in part by the National Science Foundation DMS-2323113 and IIS-2403400. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

4. Due to the space limit, we defer the experimental results in Appendix B.

References

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2012a.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012b.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. In *Conference on learning theory*, pages 76–78. PMLR, 2017.

Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 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. 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.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *International conference on machine learning*, pages 100–108. PMLR, 2014.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021.

Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration is fast and optimal. In *International Conference on Machine Learning*, pages 15239–15261. PMLR, 2023.

Tianyuan Jin, Hao-Lun Hsu, William Chang, and Pan Xu. Finite-time frequentist regret bounds of multi-agent thompson sampling on sparse hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12956–12964, 2024.

David R Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

Maximilian Kasy and Alexander Teytelboym. Matching with semi-bandits. *The Econometrics Journal*, 26(1):45–66, 2023.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161. PMLR, 2015.

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

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

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pages 223–237. PMLR, 2017.

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

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.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Yingfei Wang, Hua Ouyang, Chu Wang, Jianhui Chen, Tsvetan Asamov, and Yi Chang. Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122. PMLR, 2015.

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.

Zixin Zhong, Wang Chi Chueng, and Vincent YF Tan. Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66, 2021.

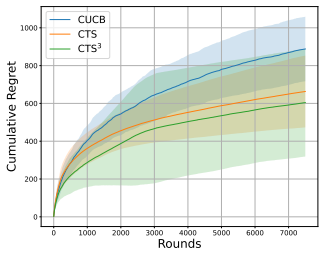
Appendix A. Related Work

Combinatorial Semi-Bandits The combinatorial multi-armed bandit (CMAB) problem was first introduced by Chen et al. (2013), which generalizes the classical multi-armed bandit (MAB) problem by allowing the agent to select a super arm consisting of multiple base arms at each round. Kveton et al. (2015b) established the first tight regret bounds for UCB-based algorithms in CMABs with linear rewards. Later, Combes et al. (2015) proposed the ESCB algorithm, which achieves improved regret bounds with an additional assumption that the feedback from selected arms is independent. Chen et al. (2016) extended the UCB-based approach to CMABs with general reward functions, providing regret bounds that depend on the smoothness of the reward function. Wang and Chen (2017) further improved the regret bounds for CMABs with linear rewards by introducing a refined analysis technique.

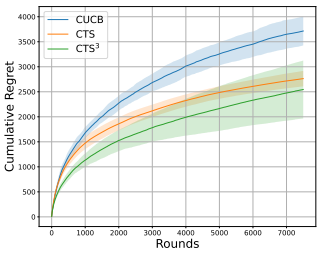
Other UCB-based algorithms have also been proposed for specific CMAB applications, such as influence maximization (Wen et al., 2015), online advertising (Wang et al., 2017), and resource allocation (Kasy and Teytelboym, 2023).

Thompson Sampling for Combinatorial Multi-Armed Bandits Thompson Sampling (TS) is a Bayesian algorithm that has been widely used in multi-armed bandit problems due to its strong empirical performance (Thompson, 1933). Agrawal and Goyal (2012b) provided the first theoretical analysis of TS for MABs with Bernoulli rewards, showing that it achieves logarithmic regret.

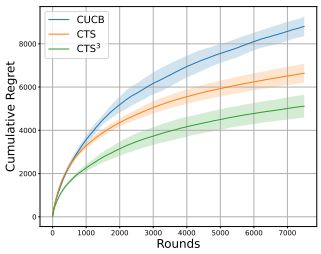
Extensions of TS to more complex bandit settings have also been studied such as combinatorial bandits. Gopalan et al. (2014) demonstrated the versatility of Thompson Sampling by establishing logarithmic regret bounds for general complex actions with non-linear rewards (such as the MAX function) and partial feedback. Komiyama et al. (2015) considered linear bandits on a uniform matroid structure (which is essentially a special case of combinatorial semi-bandits) and showed that standard TS achieves asymptotically optimal regret. Most related to our work, Wang and Chen (2018) proposed the Combinatorial Thompson Sampling (`CTS`) algorithm for CMABs with full feedback and established a regret bound of $O\left(\frac{m}{\Delta_{\min}} \log T\right) + \tilde{O}\left(\mathcal{L}^{k^*}\right)$. Recent years have witnessed several attempts to improve the regret bounds of `CTS`. Perrault et al. (2020) studied the statistical complexity of `CTS`, which provided a refined analysis specifically for the Beta-Bernoulli setting. By characterizing the posterior concentration using tight ellipsoidal confidence regions rather than independent box approximations, they effectively remove the factor m from the leading term and eliminating the exponential dependence on m in the constant term for independent Beta-Bernoulli priors. Recently, Zhang and Combes (2024) highlighted a "mismatched sampling paradox," showing that while the results by Perrault et al. (2020) hold for specific priors (like Beta), exact TS can still suffer exponential regret in general distributions. They proposed a variance-inflated Gaussian sampling technique to ensure polynomial regret for linear combinatorial bandits, further refining the conditions under which TS achieves true statistical efficiency.



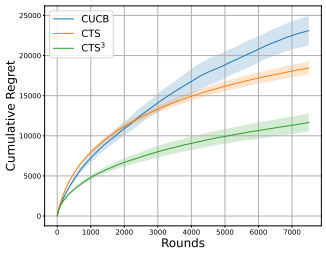
(*a*) Number of Nodes = 10



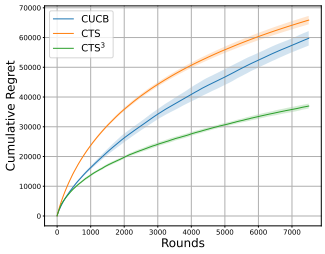
(*b*) Number of Nodes = 20



(*c*) Number of Nodes = 30



(*d*) Number of Nodes = 50



(*e*) Number of Nodes = 100

Appendix B. Experiments

To empirically validate the theoretical advantages of our proposed method, we evaluate the performance of CTS^3 against standard baselines on the Maximum Spanning Tree (MST) problem. This problem represents a canonical combinatorial setting with linear rewards and complex dependencies among the base arms (edges). We model the environment as a stochastic Maximum Spanning Tree problem on a random graph. Specifically, we generate a graph $G = (V, E)$ with $n = 10, 20, 30, 50, 100$ nodes. The edges are determined by an Erdős-Rényi model $G(n, p)$ with connection probability $p = 0.6$, resulting in a dense graph with a large number of potential super arms. For every edge $e = (u, v)$ existing in the graph, the mean weight μ_e is drawn uniformly from the interval $[0, 1]$. At each round t, the agent selects a spanning tree $S(t)$. The feedback is semi-bandit: the agent observes the noisy weight $w_{e,t} = \mu_e + \eta_{e,t}$ for every edge $e \in S(t)$, where $\eta_{e,t} \sim \mathcal{N}(0, 1)$ is independent Gaussian noise. The total reward is the sum of the true weights of the edges in the selected tree: $r(S(t), \boldsymbol{\mu}) = \sum_{e \in S(t)} \mu_e$.Algorithms Compared. We compare the following three algorithms:

- CUCB (Combinatorial UCB): The standard frequentist algorithm (Chen et al., 2013), which constructs an upper confidence bound for each edge index i as $\text{UCB}_i(t) = \hat{\mu}_i(t-1) + \sqrt{2 \ln(t)/N_i(t-1)}$.
- CTS (Standard Combinatorial Thompson Sampling): The vanilla Bayesian approach where parameter samples $\tilde{\theta}_e$ are drawn independently from the posterior distribution $\mathcal{N}(\hat{\mu}_e(t-1), 1/N_e(t-1))$ for each edge.
- CTS^3 (Ours): The proposed Single-Seed variant. Instead of independent sampling, we draw a single standard normal variable $Z \sim \mathcal{N}(0, 1)$ at each round and generate coupled samples $\tilde{\theta}_e = \hat{\mu}_e + Z \cdot \hat{\sigma}_e$. This induces perfect positive correlation (comonotonicity) among the samples to enforce concerted optimism.

As shown in the regret curves, CTS^3 demonstrates superior performance compared to the baselines. Standard CTS exhibits a steeper regret accumulation in the early-to-mid phases. This confirms our theoretical hypothesis: independent sampling fails to generate "coherent" optimistic trees, leading to inefficient exploration of the vast combinatorial space. In contrast, CTS^3 effectively synchronizes the exploration, allowing it to identify the optimal structure significantly faster. While CUCB provides a robust baseline, CTS^3 achieves lower cumulative regret over the tested horizons. This can be attributed to the Thompson Sampling property of asymptotically adapting to the problem hardness (instance-dependent bounds) better than the worst-case orientation of UCB indices.

Appendix C. Proof of Theorem 4.1

Fix an optimal super arm $S^* \in \arg\max_{S\in\mathcal{I}} r(S,\boldsymbol{\mu})$. For any super arm $S \in \mathcal{I}$, define its suboptimality gap

$$\Delta_S = r(S^*,\boldsymbol{\mu}) - r(S,\boldsymbol{\mu}) \geq 0, \qquad \Delta_{\min} = \min_{S:\Delta_S>0} \Delta_S, \qquad \Delta_{\max} = \max_{S\in\mathcal{I}} \Delta_S.$$

Let $M_\infty = \{S \in \mathcal{I} : \Delta_S = 0\}$. For each integer $r \in \mathbb{Z}$, define the gap bucket

$$M_r = \Big\{S \in \mathcal{I} : \Delta_S \in (2^{-r}, 2^{-r+1}]\Big\}.$$

Let $\mathcal{R} = \{r \in \mathbb{Z} : M_r \neq \emptyset\}$. Then $|\mathcal{R}| \leq 1 + \lceil \log_2(\Delta_{\max}/\Delta_{\min})\rceil$. The initialization phase uses at most m rounds, hence its regret is at most $m\Delta_{\max}$. In the sequel we fix a nonempty bucket M_r and bound the post-initialization regret

$$\mathbb{E}[R_{M_r}(T)] = \mathbb{E}\left[\sum_{t=t_0+1}^{T} \Delta_{S(t)}\, \mathbb{1}\{S(t) \in M_r\}\right].$$

For Thompson-sampling rounds $t \geq t_0 + 1$, let $F_{i,t}$ denote the CDF of $\mathcal{N}(\hat{\mu}_i(t-1), 1/N_i(t-1))$. Set

$$\rho_T = (Tmk)^{-8}, \qquad \theta_{\min,i}(t) = F_{i,t}^{-1}(\rho_T), \qquad \theta_{\max,i}(t) = F_{i,t}^{-1}(1-\rho_T).$$

Let $\delta_r = \frac{1}{8B}\min_{S\in M_r} \Delta_S$. Define

$$W_r(t) = \Big\{S \in M_r : \sum_{i\in S} \big(\theta_{\max,i}(t) - \mu_i\big)^+ \leq \delta_r\Big\}. \tag{C.1}$$

Since $\Delta_S \leq 16B\delta_r$ for any $S \in M_r$,

$$\begin{aligned}
\mathbb{E}[R_{M_r}(T)] &\leq 16B\delta_r \sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathbb{1}\{S(t) \in M_r\}\big] \\
&\leq \underbrace{16B\delta_r \sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathbb{1}\{S(t) \in M_r, S(t) \notin W_r(t)\}\big]}_{\heartsuit} + \underbrace{16B\delta_r \sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathbb{1}\{S(t) \in M_r, S(t) \in W_r(t)\}\big]}_{\spadesuit}.
\end{aligned}$$

Bounding term ♠. Let $c_1 = 8$ and $c_0 = 8c_1^3$. Define the dyadic grid

$$\mathcal{S}_{\mathrm{dyad}} = \{1, 2, 4, \dots, 2^{\lfloor \log_2 k \rfloor}, k\}. \tag{C.2}$$

For each dyadic $s \in \mathcal{S}_{\mathrm{dyad}}$, define

$$\tau_{r,s} = c_0 \frac{s^2 \log(Tmk)}{\delta_r^2}, \qquad \tau_r = \tau_{r,k} = c_0 \frac{k^2 \log(Tmk)}{\delta_r^2}. \tag{C.3}$$

For each round t and dyadic $s \in \mathcal{S}_{\mathrm{dyad}}$, define

$$U_{r,s}(t) = \{i \in [m] : N_i(t-1) < \tau_{r,s}\}. \tag{C.4}$$

For each $s \in [k]$, define

$$\mathrm{Good}_{i,s}^{\min}(t) = \Big\{\theta_{\min,i}(t) \geq \mu_i - \tfrac{\delta_r}{c_1 s}\Big\}, \tag{C.5}$$

$$\mathrm{Good}_{i,s}^{\max}(t) = \Big\{\theta_{\max,i}(t) \leq \mu_i + \tfrac{\delta_r}{c_1 s}\Big\}. \tag{C.6}$$

At level r, define

$$\mathcal{H}_1 = M_\infty \cup \bigcup_{a \in \mathbb{Z}: a \geq r+2} M_a, \qquad \mathcal{H}_2 = \bigcup_{a \in \mathbb{Z}: a \leq r+1} M_a.$$

Then $\mathcal{H}_1$ and $\mathcal{H}_2$ form a disjoint partition of $\mathcal{I}$. Define

$$\mathrm{Good}_{\mathcal{H}_1}^{\min}(t) = \left\{\exists S \in \mathcal{H}_1 : \sum_{i \in S} |\theta_{\min,i}(t) - \mu_i| \leq \frac{2\delta_r}{c_1}\right\}. \tag{C.7}$$

For each base arm i and integer $s \geq 1$, define

$$L_{i,s} = \mu_i - \sqrt{\frac{4 \log(Tm)}{s}}. \tag{C.8}$$

For each Thompson-sampling round $t \geq t_0 + 1$, define

$$\mathcal{E}_{1,t} = \bigcap_{s \in \mathcal{S}_{\mathrm{dyad}}} \bigcap_{i : N_i(t-1) \geq \tau_{r,s}} \big(\mathrm{Good}_{i,s}^{\min}(t) \cap \mathrm{Good}_{i,s}^{\max}(t)\big), \tag{C.9}$$

$$\mathcal{E}_{2,t} = \bigcap_{i \in [m]} \Big\{\hat{\mu}_{i,N_i(t-1)} \geq L_{i,N_i(t-1)}\Big\}. \tag{C.10}$$

Their full-horizon counterparts are

$$\mathcal{E}_1 = \bigcap_{t=t_0+1}^{T} \mathcal{E}_{1,t}, \qquad \mathcal{E}_2 = \bigcap_{t=t_0+1}^{T} \mathcal{E}_{2,t}.$$

The following lemma shows that both events hold with high probability.

Lemma C.1

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - \frac{1}{T}, \qquad \mathbb{P}(\mathcal{E}_2) \geq 1 - \frac{1}{T}.$$

Recall that in CTS^3 we draw a single standard Gaussian seed $X_t \sim \mathcal{N}(0,1)$ and generate

$$\theta_i(t) = \hat{\mu}_i(t-1) + \sqrt{\frac{1}{N_i(t-1)}} X_t, \qquad i \in [m].$$

Let Φ denote the CDF of $\mathcal{N}(0,1)$ and define

$$z_{\min} = \Phi^{-1}(\rho_T), \qquad z_{\max} = \Phi^{-1}(1-\rho_T) = -z_{\min}.$$

Define

$$\mathcal{E}_0(t) = \{X_t \in (z_{\min}, z_{\max})\} = \bigcap_{i\in[m]} \{\theta_i(t) \in (\theta_{\min,i}(t), \theta_{\max,i}(t))\}. \tag{C.11}$$

Then $\mathbb{P}(\neg\mathcal{E}_0(t)) = 2\rho_T$ for every t.

Peeling off the rare failure of $\mathcal{E}_0(t)$, $\mathcal{E}_1(t)$, and $\mathcal{E}_2(t)$. Using Lemma C.1 and a union bound, for each fixed $t \geq t_0 + 1$,

$$\mathbb{P}(\neg\mathcal{E}_0(t) \cup \neg\mathcal{E}_{1,t} \cup \neg\mathcal{E}_{2,t}) \leq 2\rho_T + \frac{1}{T} + \frac{1}{T} \leq \frac{3}{T}.$$

Therefore,

$$\begin{aligned} &16B\delta_r \sum_{t=t_0+1}^{T} \mathbb{E}\Big[\mathbb{1}\{S(t) \in M_r, S(t) \in W_r(t)\}\,\mathbb{1}\{\neg\mathcal{E}_0(t) \cup \neg\mathcal{E}_{1,t} \cup \neg\mathcal{E}_{2,t}\}\Big] \\ &\leq 16B\delta_r \sum_{t=t_0+1}^{T} \mathbb{P}(\neg\mathcal{E}_0(t) \cup \neg\mathcal{E}_{1,t} \cup \neg\mathcal{E}_{2,t}) \leq 48B\delta_r. \end{aligned} \tag{C.12}$$

Consequently,

$$\spadesuit \leq 16B\delta_r \sum_{t=t_0+1}^{T} \mathbb{E}\big[\mathbb{1}\{S(t) \in M_r, S(t) \in W_r(t), \mathcal{E}_{1,t}, \mathcal{E}_{2,t}, \mathcal{E}_0(t)\}\big] + 48B\delta_r. \tag{C.13}$$

Lemma C.2 *Fix r and a round t. Suppose the events $\mathcal{E}_0(t)$, $\mathcal{E}_{1,t}$, $\mathcal{E}_{2,t}$, and $W_r(t)$ occur. If moreover $\mathrm{Good}^{\min}_{\mathcal{H}_1}(t)$ occurs, then*

$$\max_{S\in\mathcal{H}_1} r(S, \boldsymbol{\theta}(t)) \geq \max_{S\in M_r\cap W_r(t)} r(S, \boldsymbol{\theta}(t)) + B\delta_r,$$

and hence the oracle cannot select any super arm in M_r.

Define

$$G_t = \{\neg\mathrm{Good}^{\min}_{\mathcal{H}_1}(t)\} \cap \mathcal{E}_{1,t} \cap \mathcal{E}_{2,t},$$

which is $\mathcal{F}_{t-1}$-measurable. Define the critical rounds

$$L = \{t \in \{t_0+1, \ldots, T\} : \mathbb{1}\{G_t \cap \mathcal{E}_0(t)\} = 1\}. \tag{C.14}$$

Then Lemma C.2 and (C.13) imply

$$\spadesuit \leq 16B\delta_r \mathbb{E}[|L|] + 3 \cdot 2^{-r+1}. \tag{C.15}$$

Seed-sliced classes and an under-estimation rescue event. For each t, define

$$\mathcal{E}_{\text{under}}(t) = \bigcap_{i \in S^*} \{\theta_i(t) \geq b_i\}, \qquad b_i = \mu_i - \frac{\delta_r}{c_1 k^*}. \tag{C.16}$$

For $x \in \mathbb{R}$, define

$$\theta_i(t;x) = \hat{\mu}_i(t-1) + \sqrt{\frac{1}{N_i(t-1)}}x, \qquad \boldsymbol{\theta}(t;x) = (\theta_1(t;x), \ldots, \theta_m(t;x)),$$

and let $S_t(x) = \text{ORACLE}(\boldsymbol{\theta}(t;x))$. Define

$$Y_1(t) = \{x \in \mathbb{R} : S_t(x) \in \mathcal{H}_1\}, \tag{C.17}$$
$$Y_2(t) = \{x \in \mathbb{R} : S_t(x) \in \mathcal{H}_2\}. \tag{C.18}$$

Finally define

$$L_1 = \{t \in \{t_0+1, \ldots, T\} : \mathbb{1}\{G_t \cap \mathcal{E}_0(t) \cap \mathcal{E}_{\text{under}}(t) \cap \{X_t \in Y_1(t)\}\} = 1\}, \tag{C.19}$$
$$L_2 = \{t \in \{t_0+1, \ldots, T\} : \mathbb{1}\{G_t \cap \mathcal{E}_0(t) \cap \mathcal{E}_{\text{under}}(t) \cap \{X_t \in Y_2(t)\}\} = 1\}, \tag{C.20}$$
$$L_3 = \{t \in \{t_0+1, \ldots, T\} : \mathbb{1}\{G_t \cap \mathcal{E}_0(t) \cap \neg\mathcal{E}_{\text{under}}(t)\} = 1\}. \tag{C.21}$$

Since $\mathcal{H}_1 \cup \mathcal{H}_2 = \mathcal{I}$ and $\mathcal{H}_1 \cap \mathcal{H}_2 = \emptyset$, $L = L_1 \cup L_2 \cup L_3$ is a disjoint union.

Lemma C.3 (Size of L_1 and L_2) *For the sets L_1, L_2 defined in (C.19) and (C.20),*

$$|L_1| \leq 8m\frac{\tau_r \log(2k)}{k}, \qquad |L_2| \leq 8m\frac{\tau_r \log(2k)}{k}.$$

For each round t, define

$$p_t = \mathbb{P}\left(\mathcal{E}_{\text{under}}(t) | \mathcal{F}_{t-1}, G_t, \mathcal{E}_0(t)\right). \tag{C.22}$$

Conditioned on $(\mathcal{F}_{t-1}, G_t, \mathcal{E}_0(t))$, the only remaining randomness at round t is the fresh seed X_t. A direct conditioning calculation gives

$$\begin{aligned}\mathbb{P}(G_t, \mathcal{E}_0(t), \neg\mathcal{E}_{\text{under}}(t) \mid \mathcal{F}_{t-1}) &= \mathbb{P}(G_t, \mathcal{E}_0(t) \mid \mathcal{F}_{t-1})(1-p_t) \\ &= \Big(\frac{1}{p_t} - 1\Big)\mathbb{P}(G_t, \mathcal{E}_0(t), \mathcal{E}_{\text{under}}(t) \mid \mathcal{F}_{t-1}).\end{aligned} \tag{C.23}$$

Thus

$$\mathbb{P}(G_t, \mathcal{E}_0(t) \mid \mathcal{F}_{t-1}) = \frac{1}{p_t}\mathbb{P}(G_t, \mathcal{E}_0(t), \mathcal{E}_{\text{under}}(t) \mid \mathcal{F}_{t-1}). \tag{C.24}$$

Summing over $t = t_0+1, \ldots, T$ and taking expectation yields

$$\mathbb{E}[|L|] = \mathbb{E}\Big[\sum_{t=t_0+1}^{T} \frac{\mathbb{1}\{G_t, \mathcal{E}_0(t), \mathcal{E}_{\text{under}}(t)\}}{p_t}\Big] = \mathbb{E}\Big[\sum_{t \in L_1 \cup L_2} \frac{1}{p_t}\Big]. \tag{C.25}$$

Step A: A clean tail lower bound for one coordinate. Fix t and condition on $(\mathcal{F}_{t-1}, G_t, \mathcal{E}_0(t))$. Recall $b_i = \mu_i - \delta_r/(c_1 k^*)$ and define

$$q_i(t) = \mathbb{P}(\theta_i(t) \geq b_i \mid \mathcal{F}_{t-1}, G_t, \mathcal{E}_0(t)). \tag{C.26}$$

Under CTS^3, conditionally on $\mathcal{F}_{t-1}$,

$$\theta_i(t) = \hat{\mu}_i(t-1) + \frac{1}{\sqrt{N_i(t-1)}} Z, \qquad Z \sim \mathcal{N}(0,1).$$

For any $x \geq e$, set $a(x) = \sqrt{2\log x}$. If

$$\hat{\mu}_i(t-1) \geq b_i - \sqrt{\frac{2\log x}{N_i(t-1)}}, \tag{C.27}$$

then

$$q_i(t) \geq \mathbb{P}\big(Z \geq a(x) \mid Z \in (z_{\min}, z_{\max})\big). \tag{C.28}$$

Assume $a(x) \leq z_{\max}$. Since $\mathbb{P}(Z > z_{\max}) = \rho_T$ and $\mathbb{P}(Z \in (z_{\min}, z_{\max})) = 1 - 2\rho_T$,

$$\mathbb{P}\big(Z \geq a(x) \mid Z \in (z_{\min}, z_{\max})\big) = \frac{\mathbb{P}(Z \geq a(x)) - \rho_T}{1 - 2\rho_T}. \tag{C.29}$$

By Lemma E.1, for $x \geq e$,

$$\mathbb{P}(Z \geq a(x)) \geq \frac{1}{x\sqrt{8\pi \log x}}. \tag{C.30}$$

Combining the last displays yields

$$q_i(t) \geq \frac{\frac{1}{x\sqrt{8\pi\log x}} - \rho_T}{1 - 2\rho_T}. \tag{C.31}$$

Step B: Lower bounding p_t via the single-seed property. By the single-seed identity, for any fixed super arm S,

$$\mathbb{P}\big(\forall i \in S : \theta_i(t) \geq b_i \mid \mathcal{F}_{t-1}, G_t, \mathcal{E}_0(t)\big) = \min_{i \in S} q_i(t). \tag{C.32}$$

Therefore,

$$p_t = \mathbb{P}(\mathcal{E}_{\text{under}}(t) \mid \mathcal{F}_{t-1}, G_t, \mathcal{E}_0(t)) = \min_{i \in S^*} q_i(t). \tag{C.33}$$

Step C: A uniform-in-time tail bound on $\underline{p} = \min_{t \in L} p_t$. We require the following lemma.

Lemma C.4 *Let $\hat{\mu}_s$ be the empirical mean of s random variables i.i.d. from $\mathcal{N}(\mu, 1)$. Then, for any $\beta \in \mathbb{N}^+$ and $\alpha > e$,*

$$\mathbb{P}\left(\exists s \in [\beta] : \hat{\mu}_s + \sqrt{\frac{2\log\alpha}{s}} \leq \mu\right) \leq \frac{e\log\alpha\log\beta + e + 1}{\alpha}. \tag{C.34}$$

For any $x \geq e$, define

$$\mathcal{A}(x) = \Big\{\forall i \in S^*,\ \forall s \in [\tau_r] : \hat{\mu}_{i,s} \geq \mu_i - \sqrt{\frac{2\log(k^*x)}{s}}\Big\}. \tag{C.35}$$

Lemma C.4 and a union bound over $i \in S^*$ give

$$\mathbb{P}(\mathcal{A}(x)^c) \leq \frac{e\log(k^*x)\log(\tau_r) + e + 1}{x}. \tag{C.36}$$

Let $\tilde{x} = x\log(\tau_r)$. On $\mathcal{A}(\tilde{x})$, for every $i \in S^*$ and every $t \in L$ with $N_i(t-1) = s \leq \tau_r$,

$$\hat{\mu}_i(t-1) \geq b_i - \sqrt{\frac{2\log(k^*\tilde{x})}{N_i(t-1)}}. \tag{C.37}$$

If instead $N_i(t-1) > \tau_r$, then $N_i(t-1) \geq \tau_{r,k}$ and, on $G_t \subseteq \mathcal{E}_{1,t}$, $\mathrm{Good}_{i,k}^{\min}(t)$ holds; since $k \geq k^*$, this gives $\theta_{\min,i}(t) \geq \mu_i - \delta_r/(c_1 k) \geq b_i$, and hence $q_i(t) = 1$ under $\mathcal{E}_0(t)$. Therefore, applying (C.31) with $x \leftarrow k^*\tilde{x}$ and then using (C.33), we obtain, for all $t \in L$ on $\mathcal{A}(\tilde{x})$,

$$p_t \geq \frac{\frac{1}{k^*\tilde{x}\sqrt{8\pi\log(k^*\tilde{x})}} - \rho_T}{1 - 2\rho_T}. \tag{C.38}$$

Equivalently,

$$\mathbb{P}\left(\min_{t\in L} p_t < \frac{\frac{1}{k^*x\log(\tau_r)\sqrt{8\pi\log(k^*x\log(\tau_r))}} - \rho_T}{1 - 2\rho_T}\right) \leq \frac{e\log(k^*x\log(\tau_r))\log(\tau_r) + e + 1}{x\log(\tau_r)}. \tag{C.39}$$

Step D: Bounding $\mathbb{E}[|L|]$ via a geometric peeling on $\underline{p} = \min_{t\in L} p_t$. Let $\underline{p} = \min_{t\in L} p_t$, with the convention $\underline{p} = 1$ if $L = \emptyset$. Lemma C.3 gives

$$|L_1 \cup L_2| \leq 16m\frac{\tau_r\log(2k)}{k}. \tag{C.40}$$

Combining (C.25) and (C.40),

$$\mathbb{E}[|L|] \leq 16m\frac{\tau_r\log(2k)}{k} \cdot \mathbb{E}\Big[\frac{1}{\underline{p}}\Big]. \tag{C.41}$$

On G_t we have $\mathcal{E}_{2,t}$, hence for every $t \in L$ and every $i \in [m]$,

$$\hat{\mu}_i(t-1) \geq L_{i,N_i(t-1)} = \mu_i - \sqrt{\frac{4\log(Tm)}{N_i(t-1)}}. \tag{C.42}$$

Set

$$x_{\mathrm{cap}} = (Tm)^2, \qquad a(x_{\mathrm{cap}}) = \sqrt{4\log(Tm)}. \tag{C.43}$$

Then (C.27) holds with $x = x_{\mathrm{cap}}$ for every i and $t \in L$, so

$$p_t \geq \frac{\frac{1}{x_{\mathrm{cap}}\sqrt{8\pi\log x_{\mathrm{cap}}}} - \rho_T}{1 - 2\rho_T} =: \underline{p}_{\mathrm{cap}}. \tag{C.44}$$

For Tmk below a universal constant, the regret is absorbed into the universal constant in the theorem. Hence we assume Tmk is large enough that $\underline{p}_{\text{cap}} > 0$ and $\rho_T \le \underline{p}_{\text{cap}}$. Define

$$U_{\text{sl}} = \frac{1}{\underline{p}_{\text{cap}}}. \tag{C.45}$$

Then $\underline{p} \ge 1/U_{\text{sl}}$, so the low slice contributes zero. Fix $\gamma > 1$ and define

$$Q = \left\lceil \log_\gamma(U_{\text{sl}} k^* \log(\tau_r)) \right\rceil, \qquad \kappa = \log(k^* U_{\text{sl}} \log(\tau_r)), \qquad \zeta = \sqrt{8\pi\kappa}. \tag{C.46}$$

Partition the range of $\underline{p}$ by

$$\mathcal{E}_{\text{hi}} = \left\{ \underline{p} \ge \frac{1}{k^* \log(\tau_r)} \right\}, \tag{C.47}$$

$$\mathcal{E}_s = \left\{ \underline{p} \in (\gamma^s/U_{\text{sl}}, \gamma^{s+1}/U_{\text{sl}}] \right\}, \qquad s = 0, 1, \ldots, Q-1. \tag{C.48}$$

Then

$$\mathbb{E}\Big[\frac{1}{\underline{p}}\Big] \le k^* \log(\tau_r) + \sum_{s=0}^{Q-1} \mathbb{E}\Big[\frac{1}{\underline{p}} \mathbb{1}\{\mathcal{E}_s\}\Big]. \tag{C.49}$$

For a fixed s, on $\mathcal{E}_s$,

$$\mathbb{E}\Big[\frac{1}{\underline{p}} \mathbb{1}\{\mathcal{E}_s\}\Big] \le \frac{U_{\text{sl}}}{\gamma^s} \mathbb{P}\Big(\underline{p} \le \frac{\gamma^{s+1}}{U_{\text{sl}}}\Big). \tag{C.50}$$

Define

$$x_{s+1} = \frac{U_{\text{sl}}}{2\gamma^{s+1} k^* \log(\tau_r)\zeta}. \tag{C.51}$$

Then

$$\log(k^* x_{s+1} \log(\tau_r)) = \log\left(\frac{U_{\text{sl}}}{2\gamma^{s+1}\zeta}\right) \le \kappa,$$

and hence

$$\frac{2\gamma^{s+1}}{U_{\text{sl}}} \le \frac{1}{k^* x_{s+1} \log(\tau_r) \sqrt{8\pi \log(k^* x_{s+1} \log(\tau_r))}}. \tag{C.52}$$

Since $\rho_T \le \underline{p}_{\text{cap}} = 1/U_{\text{sl}} \le \gamma^{s+1}/U_{\text{sl}}$, (C.52) implies

$$\frac{\gamma^{s+1}}{U_{\text{sl}}} \le \frac{\frac{1}{k^* x_{s+1} \log(\tau_r)\sqrt{8\pi \log(k^* x_{s+1} \log(\tau_r))}} - \rho_T}{1 - 2\rho_T}. \tag{C.53}$$

Therefore, whenever $x_{s+1} \ge e$, applying (C.39) yields

$$\begin{aligned} \mathbb{P}\Big(\underline{p} \le \frac{\gamma^{s+1}}{U_{\text{sl}}}\Big) &\le \frac{e \log(k^* x_{s+1} \log(\tau_r)) \log(\tau_r) + e + 1}{x_{s+1} \log(\tau_r)} \\ &\le \frac{e\kappa \log(\tau_r) + e + 1}{x_{s+1} \log(\tau_r)}. \end{aligned} \tag{C.54}$$

Combining (C.50), (C.51), and (C.54), for all s with $x_{s+1} \geq e$,

$$\mathbb{E}\Big[\frac{1}{\underline{p}}\,\mathbb{1}\{\mathcal{E}_s\}\Big] \leq 2\gamma k^* \log(\tau_r)\zeta\big(e\kappa\log(\tau_r) + e + 1\big). \tag{C.55}$$

If $x_{s+1} < e$, we use the trivial bound

$$\mathbb{E}\Big[\frac{1}{\underline{p}}\,\mathbb{1}\{\mathcal{E}_s\}\Big] \leq \frac{U_{\mathrm{sl}}}{\gamma^s}. \tag{C.56}$$

Let

$$s_e = \max\Big\{0, \Big\lceil \log_\gamma\Big(\frac{U_{\mathrm{sl}}}{2ek^*\log(\tau_r)\zeta}\Big)\Big\rceil - 1\Big\}. \tag{C.57}$$

Then

$$\sum_{s=s_e}^{Q-1} \mathbb{E}\Big[\frac{1}{\underline{p}}\,\mathbb{1}\{\mathcal{E}_s\}\Big] \leq \frac{2e\gamma^2}{\gamma-1}\zeta k^*\log(\tau_r). \tag{C.58}$$

Combining the slices with $\gamma = 2$ gives

$$\mathbb{E}\Big[\frac{1}{\underline{p}}\Big] = O\big(k^*\zeta Q\kappa\log^2(\tau_r)\big). \tag{C.59}$$

Together with (C.41),

$$\mathbb{E}[|L|] = O\Big(\frac{m\tau_r\log(2k)}{k}\cdot k^*\zeta Q\kappa\log^2(\tau_r)\Big). \tag{C.60}$$

Bounding term ♡. Recall

$$\heartsuit = 16B\delta_r \sum_{t=t_0+1}^{T} \mathbb{E}\big[\,\mathbb{1}\{S(t)\in M_r, S(t)\notin W_r(t)\}\big]. \tag{C.61}$$

Lemma C.5 *For each level r,*

$$\sum_{t=t_0+1}^{T} \mathbb{E}\big[\,\mathbb{1}\{S(t)\in M_r, S(t)\notin W_r(t)\}\big] \leq \frac{8m\tau_r\log(2k)}{k} + 1. \tag{C.62}$$

Consequently,

$$\heartsuit \leq 16B\delta_r\Big(8m\tau_r\log(2k)/k + 1\Big). \tag{C.63}$$

Combining the bounds for ♡ and ♠, and then summing over all nonempty levels $r \in \mathcal{R}$, yields

$$\mathbb{E}[R_{\mathcal{I}}(T)] \leq 2m\Delta_{\max} + O\Big(\frac{B^2mkk^*\mathrm{poly}(\log(T,m,\Delta_{\max}/\Delta_{\min}))}{\Delta_{\min}}\Big). \tag{C.64}$$

The worst-case bound follows by the standard gap-free conversion:

$$\begin{aligned}\mathbb{E}[R_{\mathcal{I}}(T)] &\leq 2m\Delta_{\max} + 2\max_s \min\Big\{T\delta_s/(4B), \sum_{r<s}\mathbb{E}[R_{M_r}(T)]\Big\}\\ &= 2m\Delta_{\max} + O\big(B\sqrt{mkk^*T}\cdot\mathrm{poly}(\log(T,m,\Delta_{\max}))\big).\end{aligned}$$

Appendix D. Proof of Supporting Lemmas

Proof [Proof of Lemma C.1] Recall $\rho_T = (Tmk)^{-8}$, $z_{\min} = \Phi^{-1}(\rho_T)$, and $z_{\max} = \Phi^{-1}(1-\rho_T) = -z_{\min}$. Let $Z \sim \mathcal{N}(0,1)$. By the Gaussian tail bound, $z_{\max} \le 4\sqrt{\log(Tmk)}$ for Tmk larger than a universal constant; the remaining finite cases can be absorbed into constants.

Proof of $\mathbb{P}(\mathcal{E}_1) \ge 1 - 1/T$. Fix an arm i, a sample size n, and a dyadic $s \in \mathcal{S}_{\text{dyad}}$ with $n \ge \tau_{r,s}$. Write $\varepsilon_s = \delta_r/(c_1 s)$. Since the posterior quantiles after n observations are

$$\theta_{\min,i} = \hat{\mu}_{i,n} + \frac{z_{\min}}{\sqrt{n}}, \qquad \theta_{\max,i} = \hat{\mu}_{i,n} + \frac{z_{\max}}{\sqrt{n}},$$

the choice $c_0 = 8c_1^3$ and $c_1 = 8$ ensures $z_{\max}/\sqrt{\tau_{r,s}} \le \varepsilon_s/2$. Hence, for all $n \ge \tau_{r,s}$,

$$\neg\text{Good}_{i,s}^{\min} \subseteq \{\hat{\mu}_{i,n} - \mu_i < -\varepsilon_s/2\}, \qquad \neg\text{Good}_{i,s}^{\max} \subseteq \{\hat{\mu}_{i,n} - \mu_i > \varepsilon_s/2\}.$$

Since $\hat{\mu}_{i,n} \sim \mathcal{N}(\mu_i, 1/n)$,

$$\mathbb{P}\big(\neg\text{Good}_{i,s}^{\min} \cup \neg\text{Good}_{i,s}^{\max}\big) \le \exp\Big(-\frac{n\varepsilon_s^2}{8}\Big) \le \exp\Big(-\frac{\tau_{r,s}\varepsilon_s^2}{8}\Big).$$

Moreover,

$$\tau_{r,s}\varepsilon_s^2 = \frac{c_0}{c_1^2}\log(Tmk) = 8c_1\log(Tmk),$$

and therefore the last probability is at most $(Tmk)^{-8}$. A union bound over $i \in [m]$, $n \le T$, and $s \in \mathcal{S}_{\text{dyad}}$ gives

$$\mathbb{P}(\mathcal{E}_1^c) \le mT|\mathcal{S}_{\text{dyad}}|(Tmk)^{-8} \le \frac{1}{T}.$$

Proof of $\mathbb{P}(\mathcal{E}_2) \ge 1 - 1/T$. For each i and n, $\hat{\mu}_{i,n} \sim \mathcal{N}(\mu_i, 1/n)$. Thus

$$\mathbb{P}\Big(\hat{\mu}_{i,n} < \mu_i - \sqrt{\frac{4\log(Tm)}{n}}\Big) \le \exp(-2\log(Tm)) = (Tm)^{-2}.$$

A union bound over $i \in [m]$ and $n \le T$ gives $\mathbb{P}(\mathcal{E}_2^c) \le 1/T$. ■

Proof [Proof of Lemma C.2] Fix a level r and a round t. Work on the intersection of the events stated in the lemma. There exists some $S^+ \in \mathcal{H}_1$ such that

$$\sum_{i \in S^+} |\theta_{\min,i}(t) - \mu_i| \le \frac{2\delta_r}{c_1}.$$

Since $\mathcal{E}_0(t)$ implies $\theta_i(t) \ge \theta_{\min,i}(t)$ for every i and the reward is monotone, bounded smoothness gives

$$r(S^+, \boldsymbol{\theta}(t)) \ge r(S^+, \boldsymbol{\mu}) - \frac{2B\delta_r}{c_1}. \tag{D.1}$$

Fix any $S \in M_r$. By monotonicity and bounded smoothness,

$$r(S, \boldsymbol{\theta}(t)) \le r(S, \boldsymbol{\mu}) + B\sum_{i \in S}(\theta_i(t) - \mu_i)^+.$$

On $\mathcal{E}_0(t)$, $(\theta_i(t) - \mu_i)^+ \le (\theta_{\max,i}(t) - \mu_i)^+$; on $W_r(t)$,

$$r(S, \boldsymbol{\theta}(t)) \le r(S, \boldsymbol{\mu}) + B\delta_r.$$

Since $S \in M_r$ and $\min_{S' \in M_r} \Delta_{S'} = 8B\delta_r$,

$$r(S, \boldsymbol{\theta}(t)) \le r(S^*, \boldsymbol{\mu}) - 7B\delta_r. \tag{D.2}$$

For every $S^+ \in \mathcal{H}_1$, either $S^+ \in M_\infty$ or $S^+ \in M_a$ for some $a \ge r + 2$. Hence

$$r(S^+, \boldsymbol{\mu}) \ge r(S^*, \boldsymbol{\mu}) - \frac{1}{2} \min_{S \in M_r} \Delta_S = r(S^*, \boldsymbol{\mu}) - 4B\delta_r.$$

Combining this with (D.1) and (D.2),

$$\max_{S \in \mathcal{H}_1} r(S, \boldsymbol{\theta}(t)) - \max_{S \in M_r} r(S, \boldsymbol{\theta}(t)) \ge 7B\delta_r - 4B\delta_r - \frac{2B\delta_r}{c_1} \ge B\delta_r.$$

■

Proof [Proof of Lemma C.3] Fix $j \in \{1, 2\}$. For each dyadic $s \in \mathcal{S}_{\text{dyad}}$, recall

$$U_{r,s}(t) = \{i \in [m] : N_i(t-1) < \tau_{r,s}\}, \qquad n_s(t) = |S(t) \cap U_{r,s}(t)|,$$

and define

$$\text{score}(t) = \sum_{s \in \mathcal{S}_{\text{dyad}}} \frac{2}{s} n_s(t). \tag{D.3}$$

Step 1: if $\text{score}(t) < 1$**, then** $t \notin L_1 \cup L_2$**.** We prove the contrapositive. Fix $t \in L_1 \cup L_2$ and suppose $\text{score}(t) < 1$. On $G_t \cap \mathcal{E}_0(t)$, the events $\mathcal{E}_{1,t}$ and $\mathcal{E}_{2,t}$ hold and $\theta_i(t) \in (\theta_{\min,i}(t), \theta_{\max,i}(t))$ for all i. For each arm i, define

$$s_i(t) = \max\{s \in \mathcal{S}_{\text{dyad}} : N_i(t-1) \ge \tau_{r,s}\},$$

when the set is nonempty. If this set is empty for some $i \in S(t)$, then $i \in U_{r,1}(t)$ and $\text{score}(t) \ge 2$, a contradiction. Hence $s_i(t)$ is well defined for all $i \in S(t)$. On $\mathcal{E}_{1,t}$,

$$|\theta_{\min,i}(t) - \mu_i| \le \frac{\delta_r}{c_1 s_i(t)}, \qquad |\theta_{\max,i}(t) - \mu_i| \le \frac{\delta_r}{c_1 s_i(t)}. \tag{D.4}$$

Split $S(t)$ into

$$A = \{i \in S(t) : s_i(t) = k\}, \qquad B = \{i \in S(t) : s_i(t) < k\}.$$

For $i \in B$, let $s_i^+(t)$ be the next dyadic scale after $s_i(t)$. Then $s_i^+(t) \le 2s_i(t)$, and by maximality of $s_i(t)$, $i \in U_{r,s_i^+(t)}(t)$. Thus arm i contributes at least $2/s_i^+(t) \ge 1/s_i(t)$ to $\text{score}(t)$, so

$$\sum_{i \in B} \frac{1}{s_i(t)} \le \text{score}(t). \tag{D.5}$$

Also $\sum_{i\in A} 1/k \le 1$. Therefore

$$\sum_{i\in S(t)} |\theta_{\min,i}(t) - \mu_i| \le \frac{\delta_r}{c_1}\big(1 + \mathrm{score}(t)\big) < \frac{2\delta_r}{c_1}, \tag{D.6}$$

and the same bound holds with $\theta_{\max,i}(t)$ in place of $\theta_{\min,i}(t)$.

If $t \in L_1$, then $S(t) \in \mathcal{H}_1$, and (D.6) shows that $S(t)$ witnesses $\mathrm{Good}^{\min}_{\mathcal{H}_1}(t)$, contradicting G_t. If $t \in L_2$, then $\mathcal{E}_{\mathrm{under}}(t)$ holds and $S(t) \in \mathcal{H}_2$. On $\mathcal{E}_{\mathrm{under}}(t)$,

$$r(S^*, \boldsymbol{\theta}(t)) \ge r(S^*, \boldsymbol{\mu}) - \frac{B\delta_r}{c_1}. \tag{D.7}$$

On the other hand, using monotonicity, bounded smoothness, $\mathcal{E}_0(t)$, and the upper-quantile version of (D.6),

$$r(S(t), \boldsymbol{\theta}(t)) \le r(S(t), \boldsymbol{\mu}) + \frac{2B\delta_r}{c_1}. \tag{D.8}$$

Since the oracle selects $S(t)$, (D.7) and (D.8) imply

$$\Delta_{S(t)} \le \frac{3B\delta_r}{c_1} < 2B\delta_r.$$

However $S(t) \in \mathcal{H}_2$ implies $\Delta_{S(t)} > 2^{-(r+1)} \ge 2B\delta_r$, a contradiction. Therefore every $t \in L_1 \cup L_2$ must satisfy $\mathrm{score}(t) \ge 1$.

Step 2: charge rounds by dyadic scale. For each $t \in L_j$, pick

$$s(t) \in \arg\max_{s\in\mathcal{S}_{\mathrm{dyad}}} \frac{2}{s} n_s(t), \qquad L_{j,s} = \{t \in L_j : s(t) = s\}.$$

Since $\mathrm{score}(t) \ge 1$ and $|\mathcal{S}_{\mathrm{dyad}}| \le 1 + \lfloor \log_2 k \rfloor$,

$$n_s(t) \ge \frac{s}{2|\mathcal{S}_{\mathrm{dyad}}|} \qquad \text{for all } t \in L_{j,s}. \tag{D.9}$$

Hence

$$|L_{j,s}| \le \frac{2|\mathcal{S}_{\mathrm{dyad}}|}{s} \sum_{t\in L_{j,s}} |S(t) \cap U_{r,s}(t)|.$$

For each fixed arm i, it can contribute to the last sum at most $\tau_{r,s}$ times before $N_i(\cdot)$ reaches $\tau_{r,s}$. Thus

$$|L_{j,s}| \le \frac{2|\mathcal{S}_{\mathrm{dyad}}|}{s} m\tau_{r,s}. \tag{D.10}$$

Summing over dyadic scales and using $\sum_{s\in\mathcal{S}_{\mathrm{dyad}}} s \le 2k$ gives

$$|L_j| \le 4m|\mathcal{S}_{\mathrm{dyad}}|\frac{\tau_r}{k} \le 8m\frac{\tau_r \log(2k)}{k}.$$

This proves the claim for $j = 1, 2$. ■

Proof [Proof of Lemma C.4] We directly upper bound the probability by a peeling argument. For any $\eta > 1$, define integer blocks

$$N_n = \left\lceil \frac{\beta}{\eta^{n+1}} \right\rceil, \qquad M_n = \left\lfloor \frac{\beta}{\eta^n} \right\rfloor, \qquad n = 0, 1, \dots, \lfloor \log_\eta \beta \rfloor,$$

and ignore empty blocks. These blocks, together with the singleton $s = 1$ if needed, cover $[\beta]$. By Lemma E.2,

$$\begin{aligned}
&\mathbb{P}\left(\exists s \in [\beta] : \hat{\mu}_s + \sqrt{\frac{2\log\alpha}{s}} \le \mu \right) \\
&\le \sum_{n=0}^{\lfloor \log_\eta \beta \rfloor} \mathbb{P}\left(\exists N_n \le s \le M_n : \hat{\mu}_s \le \mu - \sqrt{\frac{2\log\alpha}{M_n}} \right) + \mathbb{P}(\hat{\mu}_1 + \sqrt{2\log\alpha} \le \mu) \\
&\le \sum_{n=0}^{\lfloor \log_\eta \beta \rfloor} \exp\Big(- \frac{N_n \log\alpha}{M_n} \Big) + \frac{1}{\alpha} \le \big(\log_\eta \beta + 1 \big) \alpha^{-1/\eta} + \frac{1}{\alpha}.
\end{aligned} \tag{D.11}$$

Let $C = \log\alpha$ and choose $\eta = C/(C-1)$. Then $\eta > 1$ and $\alpha^{-1/\eta} = e/\alpha$. Moreover, since $\eta = 1 + 1/(C-1)$ and $\log(1+x) \ge x/(1+x)$ for $x > 0$,

$$\log\eta \ge \frac{1}{C}.$$

Therefore,

$$\log_\eta \beta = \frac{\log\beta}{\log\eta} \le C \log\beta = (\log\alpha)(\log\beta).$$

Combining these displays gives

$$\mathbb{P}\left(\exists s \in [\beta] : \hat{\mu}_s + \sqrt{\frac{2\log\alpha}{s}} \le \mu \right) \le \frac{e(\log\alpha)(\log\beta) + e + 1}{\alpha}.$$

■

Proof [Proof of Lemma C.5] Fix r. For each dyadic $s \in \mathcal{S}_{\mathrm{dyad}}$, define

$$U_{r,s}(t) = \{ i \in [m] : N_i(t-1) < \tau_{r,s} \}, \qquad n_s(t) = |S(t) \cap U_{r,s}(t)|,$$

and

$$\mathrm{score}(t) = \sum_{s \in \mathcal{S}_{\mathrm{dyad}}} \frac{2}{s} n_s(t). \tag{D.12}$$

On $\mathcal{E}_1$, for every t, i, and dyadic s,

$$N_i(t-1) \ge \tau_{r,s} \quad \Rightarrow \quad (\theta_{\max,i}(t) - \mu_i)^+ \le \frac{\delta_r}{c_1 s}. \tag{D.13}$$

Consider a round t such that $S(t) \in M_r$, $S(t) \notin W_r(t)$, and $\mathcal{E}_1$ holds. By definition of $W_r(t)$,

$$\sum_{i \in S(t)} (\theta_{\max,i}(t) - \mu_i)^+ > \delta_r. \tag{D.14}$$

For each arm $i \in S(t)$, define the largest reached dyadic scale

$$s_i(t) = \max\{s \in \mathcal{S}_{\text{dyad}} : N_i(t-1) \geq \tau_{r,s}\},$$

when nonempty. If this set is empty for some $i \in S(t)$, then $i \in U_{r,1}(t)$ and $\text{score}(t) \geq 2$. Otherwise, split

$$A = \{i \in S(t) : s_i(t) = k\}, \qquad B = \{i \in S(t) : s_i(t) < k\}.$$

By (D.13),

$$\sum_{i \in S(t)} (\theta_{\max,i}(t) - \mu_i)^+ \leq \frac{\delta_r}{c_1}\Big(\sum_{i \in A} \frac{1}{k} + \sum_{i \in B} \frac{1}{s_i(t)}\Big). \tag{D.15}$$

For $i \in B$, let $s_i^+(t)$ be the next dyadic scale after $s_i(t)$. Then $s_i^+(t) \leq 2s_i(t)$ and $i \in U_{r,s_i^+(t)}(t)$, so arm i contributes at least $2/s_i^+(t) \geq 1/s_i(t)$ to $\text{score}(t)$. Thus

$$\sum_{i \in B} \frac{1}{s_i(t)} \leq \text{score}(t), \qquad \sum_{i \in A} \frac{1}{k} \leq 1.$$

If $\text{score}(t) < 1$, then (D.15) is strictly smaller than δ_r, contradicting (D.14). Hence on $\mathcal{E}_1$,

$$\{S(t) \in M_r, S(t) \notin W_r(t)\} \subseteq \{\text{score}(t) \geq 1\}.$$

Define

$$\mathcal{T}_r = \{t \in \{t_0 + 1, \ldots, T\} : S(t) \in M_r, S(t) \notin W_r(t), \mathcal{E}_1\}.$$

For each $t \in \mathcal{T}_r$, choose $s(t) \in \arg\max_{s \in \mathcal{S}_{\text{dyad}}} 2n_s(t)/s$ and set $\mathcal{T}_{r,s} = \{t \in \mathcal{T}_r : s(t) = s\}$. Since $\text{score}(t) \geq 1$,

$$n_s(t) \geq \frac{s}{2|\mathcal{S}_{\text{dyad}}|} \qquad \text{for all } t \in \mathcal{T}_{r,s}.$$

Therefore,

$$|\mathcal{T}_{r,s}| \leq \frac{2|\mathcal{S}_{\text{dyad}}|}{s} \sum_{t \in \mathcal{T}_{r,s}} |S(t) \cap U_{r,s}(t)| \leq \frac{2|\mathcal{S}_{\text{dyad}}|}{s} m\tau_{r,s}.$$

Summing over s and using $\sum_{s \in \mathcal{S}_{\text{dyad}}} s \leq 2k$ gives

$$|\mathcal{T}_r| \leq 4m|\mathcal{S}_{\text{dyad}}|\frac{\tau_r}{k} \leq 8m\frac{\tau_r \log(2k)}{k}.$$

Finally,

$$\sum_{t=t_0+1}^{T} \mathbb{E}[\mathbb{1}\{S(t) \in M_r, S(t) \notin W_r(t)\}] \leq \mathbb{E}[|\mathcal{T}_r|] + \sum_{t=t_0+1}^{T} \mathbb{P}(\mathcal{E}_1^c) \leq \frac{8m\tau_r \log(2k)}{k} + 1.$$

■

Appendix E. Useful Inequalities.

Lemma E.1 (Tail Bound for Gaussian Distribution) *For a random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$,*

$$\frac{e^{-z^2/2}}{z \cdot \sqrt{2\pi}} \geq \mathbb{P}(Z > \mu + z\sigma) \geq \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-\frac{z^2}{2}}. \tag{E.1}$$

Lemma E.2 (Maximal Inequality (Lemma 4 in Ménard and Garivier (2017))) *Let N and M be two positive integers, let $\gamma > 0$, and let $\hat{\mu}_n$ be the empirical mean of n i.i.d. Gaussian random variables with mean μ and variance V. Then, for $x \leq \mu$,*

$$\mathbb{P}(\exists N \leq n \leq M, \hat{\mu}_n \leq x) \leq e^{-N(x-\mu)^2/2V}, \tag{E.2}$$

and for every $x \geq \mu$,

$$\mathbb{P}(\exists N \leq n \leq M, \hat{\mu}_n \geq x) \leq e^{-N(x-\mu)^2/(2V)}. \tag{E.3}$$