REGRET ANALYSIS OF RETRY-BASED BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We provide the first regret analysis of ReMax in stochastic multi-armed bandits. Originally introduced for reinforcement learning, ReMax is motivated by the role of exploration when multiple attempts are allowed, and is closely related to retry-based objectives such as pass@k and max@k, which value the best outcome across those attempts. Given a posterior over arm values, ReMax chooses a sampling distribution that maximizes the posterior expected maximum arm value over M virtual draws. For Gaussian rewards with K arms and any fixed integer $M \geq 2$, we characterize optimal sampling distributions through an expected-improvement balance condition and prove an $O(\sqrt{(M-1)KT\log T})$ finite-time bound on expected regret. Our analysis separates a logarithmic contribution from rounds without optimal-arm underestimation and a recovery term that dominates the bound. For every fixed instance with pairwise distinct arm means, we further prove an asymptotic logarithmic regret upper bound whose coefficient is at most $M-1$ times the optimal coefficient by Lai and Robbins. When $M = 2$, this bound implies the asymptotic optimality of ReMax. Gaussian-bandit experiments show that ReMax performs competitively with Thompson sampling and KL-UCB. They also reveal a trade-off: larger M reduces regret incurred during optimal-arm underestimation, but this reduction does not necessarily translate into lower total regret. Controlled recovery experiments show that separating suboptimal means accelerates recovery, while severe underestimation of the optimal arm can lead to long delays before it is sampled again.

1 INTRODUCTION

The stochastic multi-armed bandit problem is a fundamental framework for studying the trade-off between exploration and exploitation in sequential decision making. In the classical formulation, the learner chooses one arm per round and aims to maximize cumulative reward. Accordingly, most bandit algorithms select actions according to quantities that estimate, upper-bound, or sample from the value of a single pull. Canonical examples include optimism-based methods such as upper confidence bound (UCB) (Lai & Robbins, 1985; Auer et al., 2002; Cappé et al., 2013), empirical divergence methods (Honda & Takemura, 2010; Maillard, 2011; Bian & Jun, 2022), and Thompson sampling (Thompson, 1933; Chapelle & Li, 2011; Kaufmann et al., 2012; Komiyama et al., 2015; Agrawal & Goyal, 2017).

Retry-based objectives, which value the best outcome across multiple attempts rather than the outcome of a single attempt, provide an alternative mechanism for inducing exploration. In reinforcement learning, ReMax (Koyamada et al., 2022; Nishimori et al., 2026) instantiates this idea by optimizing the expected maximum return over multiple virtual trials under reward uncertainty. Related objectives have recently been studied in language-model reasoning, where pass@k and related max@k criteria directly optimize performance over multiple generations (Tang et al., 2025; Chen et al., 2025; Walder & Karkhanis, 2025; Bagirov et al., 2025). Improvements in pass@1 do not necessarily translate into corresponding gains in pass@k (Yue et al., 2025), further motivating objectives that explicitly account for multiple attempts. Together, these results suggest that retry-based objectives provide a promising mechanism for exploration. By studying this mechanism in a tractable bandit setting, we hope to provide theoretical insight into max@k-style optimization, including its emerging use in language-model reasoning.

Despite these empirical findings, theoretical guarantees for retry-induced exploration remain limited. Prior work has provided empirical evidence that ReMax adapts exploration to uncertainty in bandit

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.

settings and theoretical evidence that its objective induces exploration in the deterministic-reward setting (Nishimori et al., 2026; Nishimori & Parmas, 2026). Yet how this exploration behavior translates into regret guarantees for posterior-based ReMax remains theoretically unclear.

We fill this gap by providing the first regret analysis of ReMax in stochastic multi-armed bandits. For Gaussian rewards and any fixed integer $M \geq 2$, its optimal sampling distribution satisfies an expected-improvement balance condition: all arms assigned positive probability have the same posterior-averaged marginal contribution to the maximum over the other $M-1$ virtual draws, while no arm outside the support has a larger contribution. Using this characterization, we first prove an $O(\sqrt{(M-1)KT\log T})$ finite-time regret bound. The leading term comes from recovery after the optimal arm is underestimated and consequently sampled too rarely. This recovery can be particularly slow when suboptimal arms have similar empirical means, because the expected improvements between them decay slowly and can sustain their sampling probabilities. When all arm means are pairwise distinct, the expected improvement of a lower-mean suboptimal arm over a higher-mean one instead decays exponentially after sufficient sampling. Under this additional assumption, we prove an asymptotic $O(\log T)$ regret bound. Its leading coefficient is at most $M-1$ times the asymptotically optimal coefficient for general M; when $M=2$, it matches the asymptotically optimal coefficient (Lai & Robbins, 1985), establishing that ReMax is asymptotically optimal in this setting.

Contributions

1. We characterize the optimal sampling distribution of the ReMax objective through a Bayesian expected-improvement balance condition, which clarifies the structure of the ReMax sampling rule.
2. For Gaussian rewards and any fixed $M \geq 2$, we prove an $O(\sqrt{(M-1)KT\log T})$ finite-time regret bound and identify recovery from optimal-arm underestimation as the leading term. On any fixed Gaussian instance with pairwise-distinct means, we further prove a logarithmic asymptotic upper bound with coefficient at most $M-1$ times the coefficient of Lai and Robbins (1985). For $M=2$, the coefficient is exact, and ReMax is asymptotically optimal on this instance class.
3. We evaluate ReMaxGrad, a fixed-budget stochastic-gradient approximation to ReMax, on fixed and random Gaussian bandits, showing that larger M reduces underestimation regret without monotonically improving total regret. Controlled experiments with analytical $M=2$ ReMax illustrate how suboptimal-arm separation and the severity of initial underestimation affect recovery.

2 The ReMax Bandit Algorithm

2.1 Bandit Setting

We consider a K-armed stochastic bandit. We reserve μ_i for the true mean reward of arm i, $\hat{\mu}_i(t)$ for its empirical mean at round t, and θ_i for a posterior sample of arm i. At each round $t \in \{1,\dots,T\}$, the learner chooses an arm $I(t) \in [K] := \{1,\dots,K\}$ and observes a reward R_t with mean $\mu_{I(t)}$. We assume that the means are ordered as $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$, so arm 1 is uniquely optimal. For each $i > 1$, let the suboptimality gap be $\Delta_i := \mu_1 - \mu_i$ and $N_i(t) := \sum_{s=1}^{t-1} \mathbf{1}\{I(s) = i\}$, and define the (pseudo) regret as $\mathrm{Reg}(T) := \sum_{t=1}^{T}(\mu_1 - \mu_{I(t)}) = \sum_{i=2}^{K} \Delta_i N_i(T+1)$. Thus $N_i(t)$ is the number of observations of arm i available before the decision at round t. Throughout the paper, we let Π_t denote the posterior distribution over arm means after round t and write $\boldsymbol{\theta} = (\theta_1,\dots,\theta_K)$ for a sample drawn from the posterior.

2.2 Objective and Sampling Rule

Let $\pi \in \Delta^{K-1}$ be a distribution over the K arms, and let $A_{1:M} \sim \pi$ denote M i.i.d. draws from π. The ReMax objective (Nishimori et al., 2026; Koyamada et al., 2022) is

$$J^{(M)}_{\mathrm{ReMax}}(\pi;\Pi) := \mathbb{E}_{\boldsymbol{\theta}\sim\Pi}\left[\mathbb{E}_{A_{1:M}\sim\pi}\left[\max_{m\in[M]} \theta_{A_m} \,\middle|\, \boldsymbol{\theta}\right]\right]. \tag{1}$$

When the same arm appears multiple times in $A_{1:M}$, we reuse the same posterior sample θ_i for that arm. Thus, repeated occurrences of an arm correspond to repeated action selection rather than re-sampling its posterior value.

At round t, the ReMax bandit algorithm selects a distribution maximizing the ReMax objective under the current posterior:

$$\pi_t \in \arg\max_{\pi \in \Delta^{K-1}} J^{(M)}_{\mathrm{ReMax}}(\pi; \Pi_{t-1}), \qquad I(t) \sim \pi_t. \tag{2}$$

After observing the reward R_t, the posterior is updated to obtain Π_t.

2.3 Expected-Improvement Balance

A key structural property of ReMax is that, under any optimal sampling distribution, all arms in the support have the same posterior-averaged marginal improvement. For a fixed posterior sample $\boldsymbol{\theta}$, let

$$W_{M-1} := \max_{m \in [M-1]} \theta_{A_m}, \qquad A_{1:M-1} \sim \pi_t. \tag{3}$$

We define $(x)_+ := \max\{x, 0\}$ and the posterior-averaged marginal expected improvement (EI) of arm i at the optimal sampling distribution π_t by

$$\bar{s}_{t,i} := \mathbb{E}_{\boldsymbol{\theta} \sim \Pi_{t-1}} \left[\mathbb{E}_{A_{1:M-1} \sim \pi_t} \left[(\theta_i - W_{M-1})_+ \right] \right]. \tag{4}$$

Proposition 1 (Concavity and Expected-Improvement Balance)**.** *For every integer $M \geq 2$, the ReMax objective $\pi \mapsto J^{(M)}_{\mathrm{ReMax}}(\pi; \Pi_{t-1})$ is concave on Δ^{K-1}. Consequently, the simplex Karush–Kuhn–Tucker (KKT) conditions are sufficient for global optimality. Fix a round t and let π_t satisfy Equation (2). Then:*

1. *if $i, j \in \mathrm{supp}(\pi_t)$, then $\bar{s}_{t,i} = \bar{s}_{t,j}$;*
2. *if $j \in \mathrm{supp}(\pi_t)$ and $k \notin \mathrm{supp}(\pi_t)$, then $\bar{s}_{t,j} \geq \bar{s}_{t,k}$,*

where $\mathrm{supp}(\pi_t) = \{i \in [K] : \pi_{t,i} > 0\}$.

The proof follows by differentiating the ReMax objective with respect to the sampling weights and applying the KKT conditions on the simplex. See Appendix A for the full proof.

For the general best-of-M objective, draw $M-2$ additional baseline indices from π_t and write

$$W_{M-2} := \max_{m \in [M-2]} \theta_{A_m}, \qquad W_0 := -\infty.$$

Define the contextual improvement $H_{ij}(t) := \mathbb{E}_{\boldsymbol{\theta}, A_{1:M-2}} \left[\left(\theta_i - \max\{\theta_j, W_{M-2}\} \right)_+ \right]$. Splitting one of the $M-1$ baseline actions gives

$$\bar{s}_{t,i} = \sum_{j=1}^{K} \pi_{t,j} H_{ij}(t). \tag{5}$$

Thus, arms in the support equalize weighted contextual improvements. When $M = 2$, $H_{ij}(t)$ reduces to the pairwise quantity

$$G_{ij}(t) := \mathbb{E}[(\theta_i - \theta_j)_+], \qquad i \neq j,$$

and we set $G_{ii}(t) := 0$ under the duplicate-sample convention.

Discussion 1 (Difference from Conventional Expected Improvement)**.** Note that the expected-improvement quantity in Proposition 1 differs from the expected improvement algorithms studied in prior work, such as Ryzhov (2016), which are widely used in global optimization (Mockus, 1989; Vazquez & Bect, 2010; Brochu et al., 2010; Bull, 2011; Snoek et al., 2012). Proposition 1 measures how adding the M-th arm improves the maximum over the other $M-1$ arms in the same round t, whereas classical expected improvement algorithms measure how a single sample improves upon the maximum of the previous $t-1$ samples (across rounds). In fact, the analysis in Ryzhov (2016) implies that expected improvement leads to sublinear regret, although its regret guarantees are not as strong as those of other popular bandit algorithms.

2.4 Posterior-Variance Adaptation

For $M = 2$ and distinct arms i, j with Gaussian posteriors $\Pi_{t-1,k} = \mathcal{N}(\hat{\mu}_k(t), 1/N_k(t))$, let $\delta_{ij}(t) := \hat{\mu}_i(t) - \hat{\mu}_j(t)$, $\sigma^2_{ij}(t) := 1/N_i(t) + 1/N_j(t)$, and $\gamma_{ij}(t) := \delta_{ij}(t)/\sigma_{ij}(t)$. Then, by Lemma 1,

$$G_{ij}(t) = \sigma_{ij}(t)\phi(\gamma_{ij}(t)) + \delta_{ij}(t)\Phi(\gamma_{ij}(t)), \tag{6}$$

where ϕ and Φ are the standard normal density and CDF.

Equation (6) shows how posterior uncertainty directly affects expected improvement: the first term scales with the posterior standard deviation, so an uncertain arm can remain valuable even when its posterior mean is suboptimal. Figure 1 illustrates this effect for $K = M = 2$. Holding arm 1 fixed, the optimal sampling probability of arm 2 increases with its posterior variance and decreases with the posterior mean gap $\hat{\mu}_1 - \hat{\mu}_2$. Thus, ReMax naturally allocates more probability to uncertain arms whose upper tails can still improve the best-of-two objective.

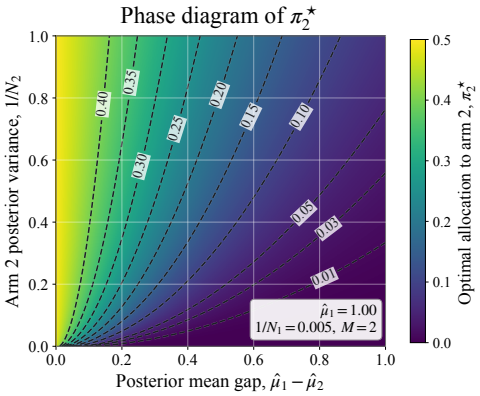

Figure 1: Optimal sampling probability $\pi_2^\star$ versus posterior mean gap and arm 2's posterior variance ($K = M = 2$).

For general M, the same upper-tail effect enters the contextual quantity $H_{ij}(t)$ in Equation (5). The relevant improvement is then measured relative to the maximum of the other virtual draws rather than a single competing arm. In contrast, Thompson sampling uses the posterior probability that an arm is optimal, whereas ReMax also accounts for the magnitude of its possible improvement over the virtual maximum.

3 Regret Analysis

We analyze the frequentist regret of ReMax for any fixed $M \geq 2$.[1] Throughout this section, expectations are taken over the reward noise and the randomness of the algorithm, with the bandit instance held fixed.

3.1 Main Results

Theorem 1 (Finite-time regret bound for ReMax)**.** *Let* $\Delta_{\max} := \max_{i \geq 2} \Delta_i$ *and* $\Delta_{\min} := \min_{i \geq 2} \Delta_i$. *Fix an integer* $M \geq 2$, *write* $r := M - 1$, *and assume* $T \geq K$. *Suppose that rewards are mutually independent across arms and pulls, with every pull of arm* i *distributed as* $\mathcal{N}(\mu_i, 1)$. *Assume that each arm is initially pulled once and that, at every subsequent round, Equation (2) is solved exactly under the joint posterior*

$$\Pi_{t-1} = \bigotimes_{i=1}^{K} \mathcal{N}\left(\hat{\mu}_i(t), \frac{1}{N_i(t)}\right).$$

For any $0 < \varepsilon \leq \min\{1, \Delta_{\min}/5\}$ *and* $D_j^\varepsilon := \frac{(\Delta_j - \frac{3}{2}\varepsilon)^2}{2}$, *there is a universal constant* $C > 0$ *such that*

$$\mathbb{E}[\mathrm{Reg}(T)] \leq \sum_{j=2}^{K} \Delta_j + \frac{C}{\varepsilon^2} \sum_{j=2}^{K} \Delta_j + 4r\Delta_{\max} \sum_{i=2}^{K} \frac{1}{\exp((\Delta_i - \varepsilon)^2/2) - 1} + \underbrace{Cr \sum_{j=2}^{K} \frac{\Delta_j}{D_j^\varepsilon} \log\left(1 + 2T\left(\frac{4}{\sqrt{\pi}\,\Delta_j}\right)^{1/r}\right)}_{\textit{Saturation}} + \underbrace{C\Delta_{\max}\sqrt{rKT\log(T+1)}\,\frac{\log(2/\varepsilon)}{\varepsilon^3}}_{\textit{Underestimation}}. \tag{7}$$

Consequently, for every fixed M *and fixed bandit instance,* $\mathbb{E}[\mathrm{Reg}(T)] = O(\sqrt{T \log T}) = o(T)$.

Theorem 1 has a logarithmic saturation term and a square-root underestimation term. Saturation is more delicate than in Thompson sampling because the EI comparison becomes effective only after both the suboptimal arm and arm 1 are sufficiently sampled. The leading term arises when an unfavorable empirical mean makes the optimal arm slow to re-enter the ReMax support. For

[1] Bayesian regret follows by integrating the frequentist bound over a prior on the arm means.

intuition, consider the $M = 2$ three-arm instance $\mu_1 > \mu_2 = \mu_3$ with $\hat{\mu}_2 = \hat{\mu}_3 = \mu_2 > \hat{\mu}_1$. Mutual EI between arms 2 and 3 decays as $t^{-1/2}$ when both counts are of order t. Balancing this quantity against the Gaussian upper-tail EI of arm 1 can delay correction of its empirical mean. The proof controls the same mechanism through the virtual-maximum EI bound and a log-count argument. Section 4.4 provides a more direct illustration of the mechanism underlying this term. Note that the regret bound of Theorem 1 is distribution-dependent as it depends on $\varepsilon = \Theta(\Delta_{\min})$, and thus we cannot directly convert it to a distribution-independent bound (Audibert & Bubeck, 2010).

The finite-time theorem allows equal suboptimal means. The preceding example indicates why a sharper conclusion becomes possible when the suboptimal means are separated: after sufficiently many samples, $\min\{G_{ij}(t), G_{ji}(t)\}$ decreases exponentially rather than at order $t^{-1/2}$. We now state the resulting pointwise asymptotic bound.

Theorem 2 (Asymptotic regret bound for ReMax)**.** *Fix an integer $M \geq 2$ and write $r := M - 1$. Under the assumptions of Theorem 1, suppose in addition that the arm means are pairwise distinct, $\mu_1 > \mu_2 > \cdots > \mu_K$. Then, for every $j > 1$,*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[N_j(T+1)]}{\log T} \leq \frac{2r}{\Delta_j^2}. \tag{8}$$

Consequently,

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \leq 2r \sum_{j=2}^{K} \frac{1}{\Delta_j}. \tag{9}$$

It is known that, for any algorithm that is uniformly good[2], the following asymptotic lower bounds hold (Lai & Robbins, 1985):

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_j(T+1)]}{\log T} \geq \frac{2}{\Delta_j^2}, \qquad j > 1. \tag{10}$$

Together with the upper bounds in Theorem 2, this shows that when $M = 2$, ReMax matches the Lai–Robbins lower bound and is therefore asymptotically optimal on every instance with pairwise-distinct arm means.

3.2 Proof Sketch

This section provides a proof sketch for the key idea for the two main theorems. One may skip this section. Write $r = M - 1$. For the finite-time bound, we separate rounds according to whether $\hat{\mu}_1(t) \leq \mu_1 - \varepsilon$, where 3ε is smaller than the smallest suboptimality gap $\min_{j>1} \Delta_j$. Similarly to the analysis of Thompson sampling (Kaufmann et al., 2012; Agrawal & Goyal, 2013), the main difficulty is that an underestimated optimal arm may receive very little sampling probability, allowing the same unfavorable estimate to persist. The proof for ReMax is much more involved because multiple posterior samples interact through the virtual maximum. The full proofs appear in Appendices B and C.

Saturation Term Consider rounds on which the optimal arm is not substantially underestimated, namely $\hat{\mu}_1(t) > \mu_1 - \varepsilon$. For a suboptimal arm j, concentration bounds the expected number of pulls with $\hat{\mu}_j(t) > \mu_j + \varepsilon$ by $2/\varepsilon^2$. On the remaining rounds, $\hat{\mu}_1(t) - \hat{\mu}_j(t) > \Delta_j - 2\varepsilon > 0$, so the empirical means correctly rank arm 1 above arm j.

The key point is that this empirical gap makes arm j unlikely to be selected once both arms have enough observations. For each j, choose a sample threshold $L_j = O(\log T)$ large enough that the EI balance condition gives $\pi_{t,j} \leq 1/(2T)$ whenever $N_1(t), N_j(t) \geq L_j$ on these rounds. We then count pulls of arm j in three cases. Before $N_j(t)$ reaches L_j, there are at most L_j such pulls. If $N_j(t) \geq L_j$ but $N_1(t) < L_j$, the EI comparison gives $\pi_{t,j} \leq \pi_{t,1}$. Thus, the expected number of pulls of arm j in this case is at most the expected number of pulls of arm 1 in the same case, which is at most L_j. Finally, when both counts are at least L_j, summing $\pi_{t,j} \leq 1/(2T)$ over at most T rounds gives at most $1/2$ expected pulls. Combining these cases with the overestimation bound, and weighting each arm's pull count by its gap Δ_j, yields the logarithmic term in Theorem 1.

[2]An algorithm is uniformly good if its regret is $o(T^a)$ for all $a > 0$ and all instances.

Underestimation Term As in the analysis of Thompson sampling, underestimation of the optimal arm requires more detailed analysis of the sampling dynamics than overestimation. Group the underestimated rounds into blocks with $N_1(t) = n$. Within such a block, the empirical mean $\hat{\mu}_{1,n}$ of the first n rewards from arm 1 and its posterior remain unchanged. Let e_n denote its posterior expected improvement above $\mu_1 - \varepsilon$.

We first separate rounds on which sufficiently sampled but empirically optimistic suboptimal arms carry substantial sampling probability. Their expected number is controlled by charging them to upward deviations of suboptimal empirical means. On the remaining rounds, either substantial probability lies on insufficiently sampled arms, or EI balance forces the sampled arms to retain enough posterior uncertainty. In both cases, the log-count potential $\Psi_t := \sum_{i=1}^{K} \log N_i(t)$ has conditional expected increment at least ce_n^2/r, for a universal constant $c > 0$. Since its total increase is at most $K \log(T+1)$, the conditional expected number of remaining rounds in the count-n block is at most B_T/e_n^2, where $B_T = CrK \log(T+1)$ for a universal constant $C > 0$. Combining this with the horizon bound gives

$$\min\left\{T, \frac{B_T}{e_n^2}\right\} \le \frac{\sqrt{B_T T}}{e_n}.$$

This step produces the square-root dependence on T. Averaging over $\hat{\mu}_{1,n} \sim \mathcal{N}(\mu_1, 1/n)$ and summing over n bounds the expected number of these rounds by

$$C\sqrt{rKT \log(T+1)} \frac{\log(2/\varepsilon)}{\varepsilon^3}.$$

Multiplying the round-count bounds by $\Delta_{\max}$ gives the underestimation contribution in Theorem 1.

Asymptotic Regret For Theorem 2, pairwise separation allows a sharper recovery argument. Fix $b \in (\max_{i>1} \mu_i, \mu_1)$. After sufficiently many observations, the expected improvement of a lower-mean suboptimal arm over a higher-mean one decays exponentially in their smaller sample count. Choose sampling thresholds so that these improvements are small relative to the optimal arm's EI above b. EI balance then forces some sampling probability onto arm 1 or a suboptimal arm that has not yet reached its threshold. Each such draw either ends the current count-n block or moves a suboptimal arm closer to its threshold.

The resulting waiting-time bound involves $1/\mathrm{EI}$, up to polynomial factors, where EI is the optimal arm's expected improvement above b. To see why this bound is integrable, write $\varepsilon_b := \mu_1 - b > 0$ and $\hat{\mu}_{1,n} = b - y$ with $y \ge 0$. The Gaussian EI bound gives

$$\frac{1}{\mathrm{EI}} \le C\sqrt{n}(1 + ny^2)e^{ny^2/2}.$$

Meanwhile, the density of $\hat{\mu}_{1,n} \sim \mathcal{N}(\mu_1, 1/n)$ at $b - y$ is proportional to $\sqrt{n}\, e^{-n(\varepsilon_b + y)^2/2}$. Thus, when averaging the waiting-time bound, the exponential factors combine as

$$e^{ny^2/2} e^{-n(\varepsilon_b + y)^2/2} = e^{-n\varepsilon_b^2/2} e^{-n\varepsilon_b y}.$$

The remaining decay in y makes the polynomial factors integrable. Including these factors and integrating over $y \ge 0$ bounds the expected contribution from the block with $N_1(t) = n$ by $C_b n^2 e^{-n\varepsilon_b^2/2}$, where C_b is independent of n and the horizon. Summing over n therefore gives

$$\mathbb{E}\left[\sum_{t=K+1}^{\infty} \mathbf{1}\{\hat{\mu}_1(t) \le b\}\right] \le C_b \sum_{n=1}^{\infty} n^2 e^{-n\varepsilon_b^2/2} < \infty. \tag{11}$$

Thus the expected underestimation contribution is bounded independently of the horizon.

To obtain the leading logarithmic coefficient of the saturation term, we need a sharper comparison than that of Theorem 1 because we must optimize the constant factor. The analysis here is more challenging than Thompson sampling because we must lower-bound $N_1(t)$ unlike in Thompson sampling, where only the suboptimal arm's count needs to be controlled. Intuitively, the optimal arm should receive far more observations than a suboptimal arm, so that $N_1(t) \gg N_j(t)$: Its posterior uncertainty is then negligible relative to that of arm j, and the effective sample count in the pairwise EI bound satisfies

$$\frac{N_1(t)N_j(t)}{N_1(t) + N_j(t)} = \frac{N_j(t)}{1 + N_j(t)/N_1(t)} \approx N_j(t).$$

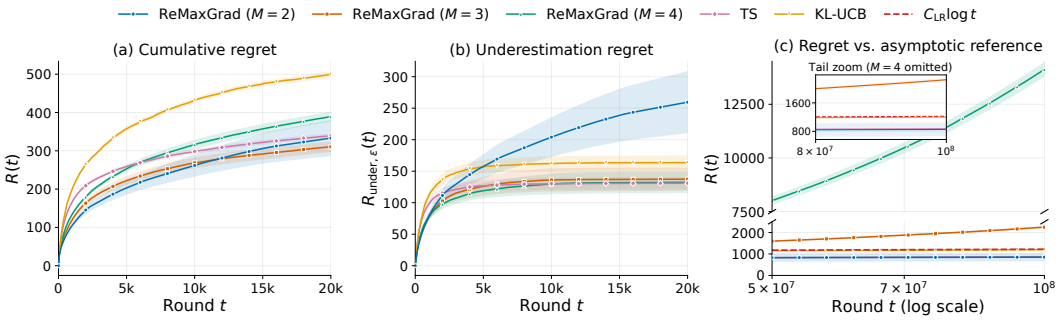

Figure 2: Regret means over 100 trajectories with one-standard-error bands. The red dashed line is $C_{\mathrm{LR}} \log t$, with $C_{\mathrm{LR}} \approx 66.36$. (a) Cumulative and (b) underestimation regret through $t = 20{,}000$. (c) Regret for $5 \times 10^7 \le t \le 10^8$ on a log time axis; the inset shows $8 \times 10^7 \le t \le 10^8$ without $M = 4$. In (c), TS largely overlaps ReMaxGrad ($M = 2$).

Thus the leading exploration cost is determined by the observations needed to distinguish arm j from an essentially known optimal mean. However, we need to carefully control the rounds on which this is not yet the case. The proof makes this intuition precise by controlling separately the rounds on which the count ratio is not large. Fix $j > 1$, sufficiently small $\varepsilon > 0$, and $\eta > 0$, and set

$$d_{j,\varepsilon} := \frac{(\Delta_j - 2\varepsilon)^2}{2}, \qquad m_T := \left\lceil \frac{r(1+\eta)\log T}{d_{j,\varepsilon}} \right\rceil,$$

where $\eta > 0$ is a small constant. On rounds with $\hat{\mu}_1(t) > \mu_1 - \varepsilon$ and $\hat{\mu}_j(t) \le \mu_j + \varepsilon$, the saturation inequality gives

$$\pi_{t,j} \le A \exp\left\{ -\frac{d_{j,\varepsilon}}{r} \frac{N_1(t) N_j(t)}{N_1(t) + N_j(t)} \right\},$$

where $A > 0$ is independent of t and T. There are at most m_T pulls before arm j reaches this threshold. Optimal-arm underestimation and empirical overestimation of arm j contribute only $O(1)$ expected pulls.

For the remaining rounds, choose a sufficiently large fixed κ such that $(1+\eta)\kappa/(\kappa+1) > 1$, and split according to whether $N_1(t) \ge \kappa N_j(t)$. In this case, the effective sample count is at least $\kappa N_j(t)/(\kappa+1)$, so after $N_j(t) \ge m_T$ the sampling probability is at most $AT^{-(1+\eta)\kappa/(\kappa+1)}$. Summing over at most T rounds gives an $o(1)$ contribution. In the complementary case, group rounds by $N_1(t) = n$. The proof bounds the lengths of non-underestimated blocks by a common integrable random bound; combining this with the sampling-probability bound gives an $O(1)$ contribution after saturation. Thus, rounds on which $N_1(t)$ is not much larger than $N_j(t)$ do not change the leading logarithmic coefficient. Hence

$$\mathbb{E}[N_j(T+1)] \le m_T + O(1).$$

Dividing by $\log T$, then letting $\eta \downarrow 0$ and $\varepsilon \downarrow 0$, gives the coefficient $2r/\Delta_j^2$. Summing over suboptimal arms yields the regret bound. When $M = 2$, the upper bound matches the Lai–Robbins lower bound, giving the stated limits and asymptotic optimality.

4 Experiments

We evaluate ReMax on fixed and randomly generated Gaussian bandits and examine recovery from optimal-arm underestimation in two controlled three-arm experiments. The recovery experiments vary the separation between suboptimal means and the severity of the initial underestimation. We use ReMaxGrad, a stochastic-gradient approximation to the ReMax optimizer, with retry budgets $M \in \{2, 3, 4\}$ for the regret comparisons, and the analytical $M = 2$ solution for the recovery experiments. Appendix D provides implementation details and parameter settings.

4.1 Experimental protocol

The fixed- and random-instance experiments use independent rewards from $\mathcal{N}(\mu_i, 1)$, and each method begins by pulling every arm once. We compare ReMaxGrad (Nishimori et al., 2026)

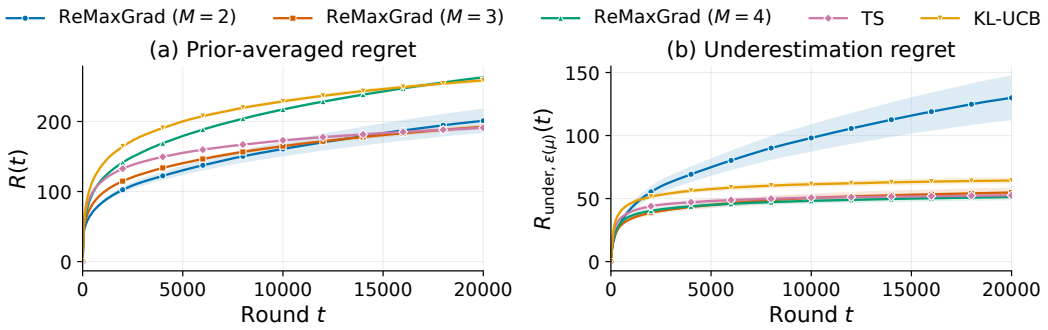

Figure 3: Means over 1,000 random $K = 16$ instances ($\mu_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$), with one-standard-error bands. (a) Prior-averaged cumulative regret; (b) underestimation regret through $t = 20{,}000$.

with Gaussian Thompson sampling (TS) and Gaussian KL–UCB. ReMaxGrad and TS use the flat-prior Gaussian posterior $\Pi_{t-1,i} = \mathcal{N}(\hat{\mu}_i(t), 1/N_i(t))$, whereas Gaussian KL–UCB selects an arm maximizing the index $\hat{\mu}_i(t) + \sqrt{2\log(t-1)/N_i(t)}$.

For each instance, let $i^\star \in \arg\max_i \mu_i$ and $\mu_\star = \mu_{i^\star}$. We report cumulative pseudo-regret and the regret incurred during optimal-arm underestimation after initialization:

$$\mathrm{Reg}(t) := \sum_{s=1}^{t} \big(\mu_\star - \mu_{I(s)}\big), \qquad \mathrm{Reg}_{\mathrm{under},\varepsilon}(t) := \sum_{s=K+1}^{t} \big(\mu_\star - \mu_{I(s)}\big)\mathbf{1}\{\hat{\mu}_{i^\star}(s) < \mu_\star - \varepsilon\}\,.$$

The second metric isolates the contribution of optimal-arm underestimation to cumulative regret.

4.2 Fixed frequentist instance

We construct a fixed $K = 16$ Gaussian instance with evenly spaced means, using $g = 0.1$: $\mu_i = 1.5 - g(i-1)$, $i \in [16]$. The means are pairwise distinct, as required by Theorem 2. Underestimation regret uses the threshold $\varepsilon = 0.02$. We average over 100 trajectories, each initialized with an independent seed, and continue the same trajectories to $T = 10^8$.

Results Figure 2 (a) shows that $M = 3$ attains the lowest mean cumulative regret at $t = 20{,}000$, while $M = 2$ remains close to TS. Figure 2 (b) shows that increasing M reduces underestimation regret, although this reduction does not translate into a monotonic improvement in total regret. At large horizons, Figure 2 (c) shows that $M = 2$ has a slope comparable to those of TS and KL–UCB on the logarithmic time axis, while the slopes become progressively steeper for $M = 3$ and $M = 4$. The observed behavior is consistent with the asymptotic result in Theorem 2.

4.3 Random-instance experiment

We independently sample 1,000 mean vectors for random $K = 16$ instances according to $\mu_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, where $i \in [16]$. Each mean vector defines one instance, on which every method is run once. Underestimation regret uses the instance-specific threshold $\varepsilon = \min\{1, \Delta_{\min}/5\}$, where $\Delta_{\min} := \min_{i \neq i^\star}(\mu_\star - \mu_i)$. Averaging over instances and stochastic trajectories yields the prior-averaged regret.

Results Figure 3 shows that $M = 2$ and $M = 3$ have mean cumulative regret comparable to TS and lower than KL–UCB at $t = 20{,}000$. The $M = 4$ variant performs similarly to KL–UCB. As in the fixed instance, increasing M reduces underestimation regret without monotonically improving total regret.

4.4 Three-arm recovery from underestimation

We first examine how suboptimal-arm separation affects recovery. Consider the three-arm family $\boldsymbol{\mu}(\delta) = (1.5, 1, 1-\delta)$ for $\delta \in \{0, 0.05, 0.1, 0.2\}$, with unit reward variance. Arm 1 has a constant gap of 0.5 to the best suboptimal arm, while arms 2 and 3 are tied at $\delta = 0$ and separated for $\delta > 0$.

We prescribe an initial state with 16 observations per arm and empirical means $(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (0.8, 1, 1-\delta)$. Subsequent rewards and posterior updates follow the usual bandit dynamics, with flat-prior Gaussian posteriors. For each δ, we simulate 10,000 independently seeded trajectories and record $\tau_{\text{return}} := \inf\{s \geq 1 : I(s) = 1\}$, the first return to arm 1 after initialization.

Results Figure 4 shows the slowest recovery when the suboptimal arms are tied. Increasing their separation leads to progressively faster recovery. This illustrates how competition between suboptimal arms can prolong optimal-arm underestimation, and helps explain the role of pairwise separation in Theorem 2.

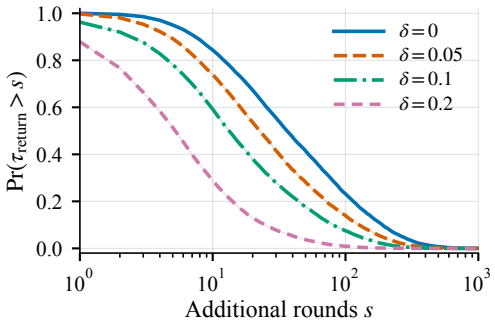

Figure 4: Curves are estimated from 10,000 trajectories per δ, with shaded bands denoting one standard error. Each curve shows the probability that arm 1 remains unselected after s additional rounds.

4.5 Empirical Evidence for Finite Waiting Times

To empirically verify the finiteness of the underestimation term (i.e., (11)), we vary the severity of the initial underestimation. Here the three arms have means $(1.0, 0.9, 0.8)$ and reward standard deviation $\sigma = 0.1$. Each arm is initially pulled once, and arm 1's first observation is replaced by $x(q) := 1 + \sigma\Phi^{-1}(q)$, where Φ is the standard normal CDF. Thus, q is the worst q-quantile of the optimal arm's reward distribution. For example, $q = 0.01$ gives $x(q) \approx 0.767$, an observation in the bottom 1% of the reward distribution. All other observations are sampled normally. We use the flat-prior posterior $\mathcal{N}(\hat{\mu}_i, \sigma^2/N_i)$.

We aim to measure the worst-case additional waiting time for the optimal arm to be revisited. To see this, let $\tau_{\text{return}} := \inf\{s \geq 1 : I(s) = 1\}$ be the number of rounds until the algorithm draws arm 1 again. Let $m(q) := \mathbb{E}[\tau_{\text{return}} \mid X_{1,1} = x(q)]$. To have a finite expected waiting time, the second draw of the optimal arm[3] is required to occur quickly enough on average, which is captured by:

$$\int_0^1 \underbrace{m(q)}_{\text{expected waiting time given at the worst } q\text{-quantile}} dq < \infty, \tag{12}$$

and for this equation to hold, $m(q)$ must grow slower than $1/q$ as $q \to 0$, as $\int (1/q)dq = \infty$.

To measure $m(q)$ empirically, we use 31 logarithmically spaced values of q from 10^{-2} to 10^{-7} and 1,000 independently seeded trajectories per value, reusing the same seeds across values of q. Every trajectory returns before the cap of 10^8 additional rounds.

Results Figure 5 shows $m(q)$ as a function of $1/q$. We observe that $m(q)$ grows slower than $1/q$, empirically supporting Eq. (12). At the same time, the mean recovery time increases from tens to millions of rounds as the initial observation moves deeper into the lower tail, approximately following the $1/q$ scale over the tested range. Thus, even with pairwise-distinct means, severe initial underestimation can substantially delay the return to the optimal arm.

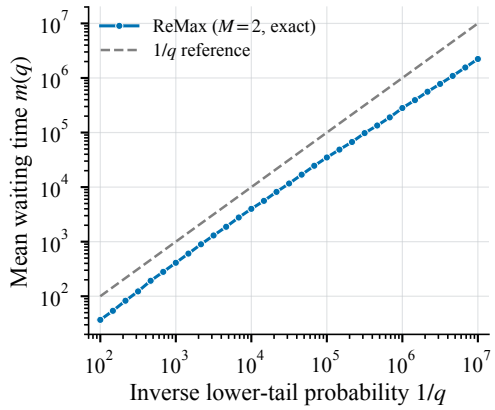

Figure 5: Mean $m(q)$ over 1,000 trajectories per q, with one-standard-error bands, plotted against $1/q$; the dashed line shows the $1/q$ reference.

[3] One can also show that this quantity diverges for a UCB algorithm with insufficient exploration, where the exploration term is smaller than $\sqrt{2\sigma^2 \log t / N_i(t)}$. Therefore, the finiteness of this quantity supports the sufficiency of exploration.

AI USE STATEMENT

We used generative AI tools to suggest the approach underlying the asymptotic-optimality result in Theorem 2 and to assist in developing its proof. We also used generative AI tools to check and improve mathematical arguments, assist with experimental code, and edit and polish the manuscript. The AI-assisted proofs were subsequently reviewed line by line by multiple authors, who independently checked their correctness and revised them to improve clarity and readability. All AI-assisted code was also reviewed and verified by the authors.

ETHICS STATEMENT

We are not aware of any ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

All assumptions and theoretical results are stated in the main text, with complete proofs provided in the appendices. Detailed descriptions of the experimental setup, implementation of ReMaxGrad, hyperparameters, and additional experimental results are also provided in the appendices.

REFERENCES

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson Sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, JMLR Workshop and Conference Proceedings, pp. 99–107. JMLR.org, 2013. URL http://proceedings.mlr.press/v31/agrawal13a.html.

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

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11:2785–2836, 2010. doi: 10.5555/1756006.1953023. URL https://dl.acm.org/doi/10.5555/1756006.1953023.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL http://dx.doi.org/10.1023/A:1013689704352.

Farid Bagirov, Mikhail Arkhipov, Ksenia Sycheva, Evgeniy Glukhov, and Egor Bogomolov. The best of n worlds: Aligning reinforcement learning with best-of-n sampling via max@ k optimisation. *arXiv preprint arXiv:2510.23393*, 2025.

Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, Proceedings of Machine Learning Research, pp. 54–72. PMLR, 2022. URL https://proceedings.mlr.press/v151/bian22a.html.

Eric Brochu, Vlad M Cora, and Nando de Freitas. A tutorial on Bayesian Optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

Adam D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(88):2879–2904, 2011. URL http://jmlr.org/papers/v12/bull11a.html.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 06 2013. doi: 10.1214/13-AOS1119. URL http://dx.doi.org/10.1214/13-AOS1119.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson Sampling. *Advances in Neural Information Processing Systems*, 24, 2011.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In Adam Tauman Kalai and Mehryar Mohri (eds.), *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pp. 67–79. Omnipress, 2010. URL http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=75.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An asymptotically optimal finite-time analysis. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann (eds.), *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, Lecture Notes in Computer Science, pp. 199–213. Springer, 2012. doi: 10.1007/978-3-642-34106-9_18. URL https://doi.org/10.1007/978-3-642-34106-9_18.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of Thompson Sampling in stochastic multi-armed bandit problem with multiple plays. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1152–1161, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/komiyama15.html.

Sotetsu Koyamada, Paavo Parmas, Tadashi Kozuno, and Shin Ishii. Emergence of exploration in policy gradient reinforcement learning via resetting. 2022. URL https://openreview.net/forum?id=GKsNIC_mQRG.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Odalric-Ambrym Maillard. *SEQUENTIAL LEARNING: Bandits, Statistics and Reinforcement.* PhD thesis, University of Science and Technology of Lille - Lille I, October 2011. URL https://theses.hal.science/tel-00845410.

Jonas Mockus. *Bayesian Approach to Global Optimization*, volume 37 of *Mathematics and Its Applications*. Kluwer Academic Publishers, 1989.

Soichiro Nishimori and Paavo Parmas. Retry policy gradients in continuous action spaces. *arXiv preprint arXiv:2606.05888*, 2026.

Soichiro Nishimori, Paavo Parmas, Sotetsu Koyamada, Tadashi Kozuno, Toshinori Kitamura, Shin Ishii, and Yutaka Matsuo. Emergence of exploration in policy gradient reinforcement learning via retrying. In *Proceedings of the International Conference on Machine Learning*, 2026.

Ilya O. Ryzhov. On the convergence rates of expected improvement methods. *Oper. Res.*, 64(6): 1515–1528, December 2016. ISSN 0030-364X. doi: 10.1287/opre.2016.1494. URL https://doi.org/10.1287/opre.2016.1494.

Jasper Snoek, Hugo Larochelle, and Ryan Adams. Practical Bayesian Optimization of machine learning algorithms. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Remi Munos. Optimizing language models for inference time objectives using reinforcement learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59066–59085. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tang25o.html.

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.

Lieven Vandenberghe and Stephen Boyd. *Convex optimization*, volume 1. Cambridge university press Cambridge, 2004.

Emmanuel Vazquez and Julien Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11): 3088–3095, 2010.

Christian Walder and Deep Tejas Karkhanis. Pass@k policy optimization: Solving harder reinforcement learning problems. *Advances in Neural Information Processing Systems*, 38:152416–152445, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2025.

A PROOF OF PROPOSITION 1

Proof of Proposition 1. Fix a round t and abbreviate

$$J(\pi) := J^{(M)}_{\mathrm{ReMax}}(\pi; \Pi_{t-1}).$$

For a fixed posterior sample θ, define

$$v(\pi \mid \theta) := \mathbb{E}_{A_{1:M}\sim\pi}\left[\max_{m\in[M]} \theta_{A_m}\right].$$

We first prove concavity. Let

$$x_1 < \cdots < x_L$$

be the distinct values among $\theta_1, \ldots, \theta_K$, and define

$$Q_\ell(\pi) := \sum_{i:\, \theta_i \le x_\ell} \pi_i, \qquad \ell \in [L].$$

The maximum of the M draws is at most x_ℓ with probability $Q_\ell(\pi)^M$. Therefore,

$$\begin{aligned} v(\pi \mid \theta) &= x_1 + \sum_{\ell=1}^{L-1} (x_{\ell+1} - x_\ell)\big(1 - Q_\ell(\pi)^M\big) \\ &= x_L - \sum_{\ell=1}^{L-1} (x_{\ell+1} - x_\ell) Q_\ell(\pi)^M. \end{aligned}$$

Each Q_ℓ is affine on the simplex, and $q \mapsto q^M$ is convex on $[0,1]$. Hence $v(\cdot \mid \theta)$ is concave on Δ_{K-1}. Taking expectations over $\theta \sim \Pi_{t-1}$ shows that J is concave.

Expanding over all realizations of $a_1, \ldots, a_M \in [K]$ gives

$$v(\pi \mid \theta) = \sum_{a_1=1}^{K} \cdots \sum_{a_M=1}^{K} \left(\prod_{m=1}^{M} \pi_{a_m}\right) \max\{\theta_{a_1}, \ldots, \theta_{a_M}\}.$$

The Gaussian posterior satisfies

$$\mathbb{E}_{\theta\sim\Pi_{t-1}}\left[\max_{i\in[K]} |\theta_i|\right] < \infty.$$

Thus, after averaging the finite expansion above, J is a polynomial in π with finite coefficients. In particular, its partial derivatives exist at both interior and boundary points of the simplex, and differentiation may be passed through the posterior expectation.

By the product rule,

$$\frac{\partial}{\partial \pi_i}\left(\prod_{m=1}^{M} \pi_{a_m}\right) = \sum_{k=1}^{M} \mathbf{1}\{a_k = i\} \prod_{m\neq k} \pi_{a_m}.$$

Substituting this identity into the expansion gives

$$\frac{\partial v(\pi \mid \theta)}{\partial \pi_i} = \sum_{a_1,\ldots,a_M} \left(\sum_{k=1}^{M} \mathbf{1}\{a_k = i\} \prod_{m\neq k} \pi_{a_m}\right) \max\{\theta_{a_1}, \ldots, \theta_{a_M}\} \tag{13}$$

$$= \sum_{k=1}^{M} \sum_{a_1,\ldots,a_M} \mathbf{1}\{a_k = i\} \left(\prod_{m\neq k} \pi_{a_m}\right) \max\{\theta_{a_1}, \ldots, \theta_{a_M}\}. \tag{14}$$

By exchangeability, the M terms in the outer sum are identical. For any fixed position k, the condition $a_k = i$ fixes that draw to arm i, while the remaining $M-1$ indices are independently drawn from π. Writing

$$W_{M-1} := \max_{m\in[M-1]} \theta_{A_m}, \qquad A_{1:M-1} \sim \pi,$$

we obtain

$$\frac{\partial v(\pi \mid \theta)}{\partial \pi_i} = M\,\mathbb{E}_{A_{1:M-1}\sim\pi}\left[\max\{\theta_i, W_{M-1}\}\right].$$

Define

$$s_{t,i}(\pi,\theta) := \mathbb{E}_{A_{1:M-1}\sim\pi}\left[(\theta_i - W_{M-1})_+\right].$$

Using

$$\max\{\theta_i, W_{M-1}\} = (\theta_i - W_{M-1})_+ + W_{M-1},$$

we have

$$\frac{\partial v(\pi \mid \theta)}{\partial \pi_i} = M\left(\mathbb{E}_{A_{1:M-1}\sim\pi}\left[(\theta_i - W_{M-1})_+\right] + \mathbb{E}_{A_{1:M-1}\sim\pi}[W_{M-1}]\right) \tag{15}$$

$$= M\left(s_{t,i}(\pi,\theta) + \mathbb{E}_{A_{1:M-1}\sim\pi}[W_{M-1}]\right). \tag{16}$$

Let

$$C_t(\pi) := \mathbb{E}_{\theta\sim\Pi_{t-1}}\left[\mathbb{E}_{A_{1:M-1}\sim\pi}[W_{M-1}]\right].$$

Averaging over $\theta \sim \Pi_{t-1}$ gives

$$\frac{\partial J(\pi)}{\partial \pi_i} = \mathbb{E}_{\theta\sim\Pi_{t-1}}\left[\frac{\partial v(\pi \mid \theta)}{\partial \pi_i}\right] \tag{17}$$

$$= M\left(\mathbb{E}_{\theta\sim\Pi_{t-1}}[s_{t,i}(\pi,\theta)] + C_t(\pi)\right). \tag{18}$$

At the optimizer π_t,

$$\frac{\partial J(\pi_t)}{\partial \pi_i} = M\left(\bar{s}_{t,i} + C_t(\pi_t)\right), \qquad \bar{s}_{t,i} := \mathbb{E}_{\theta\sim\Pi_{t-1}}[s_{t,i}(\pi_t,\theta)].$$

Since J is differentiable and concave, the simplex KKT conditions are sufficient for global optimality. At π_t, there exists $\lambda_t \in \mathbb{R}$ such that

$$\frac{\partial J(\pi_t)}{\partial \pi_i} = \lambda_t \quad \text{for } i \in \operatorname{supp}(\pi_t),$$

and

$$\frac{\partial J(\pi_t)}{\partial \pi_k} \le \lambda_t \quad \text{for } k \notin \operatorname{supp}(\pi_t).$$

Since $C_t(\pi_t)$ is common to all arm indices, these relations imply

$$\bar{s}_{t,i} = \bar{s}_{t,j} \quad \text{for } i, j \in \operatorname{supp}(\pi_t),$$

and

$$\bar{s}_{t,j} \ge \bar{s}_{t,k} \quad \text{for } j \in \operatorname{supp}(\pi_t), \quad k \notin \operatorname{supp}(\pi_t).$$

This proves the claims. □

B Proof of Theorem 1

B.1 Setup and decomposition

Throughout this appendix, fix the setting of Theorem 1 and write

$$r := M - 1.$$

For each arm i, write its reward sequence as

$$X_{i,m} \sim \mathcal{N}(\mu_i, 1), \qquad \hat{\mu}_{i,m} := \frac{1}{m}\sum_{s=1}^{m} X_{i,s}, \qquad m \ge 1. \tag{19}$$

The variables $X_{i,m}$ are mutually independent over all pairs (i,m) and are independent of the fresh random variables used by the algorithm to draw its actions. The mth pull of arm i reveals $X_{i,m}$, so the pathwise identity $\hat{\mu}_i(t) = \hat{\mu}_{i,m}$ holds whenever $N_i(t) = m$.

Independently of the adaptive sampling rule, the fixed-table average has the unconditional distribution $\hat{\mu}_{i,m} \sim \mathcal{N}(\mu_i, 1/m)$. Let $\mathcal{F}_0$ be trivial and, for $t \geq 1$, define

$$\mathcal{F}_t := \sigma(I(s), R_s : 1 \leq s \leq t)\,.$$

Thus, $\mathcal{F}_{t-1}$ is simply the interaction history available before the action at round t is chosen. We use the pre-decision convention: $N_i(t)$, $\hat{\mu}_i(t)$, the posterior Π_{t-1}, and the sampling distribution π_t are all determined by this history. Since the algorithm then draws $I(t)$ from π_t using fresh randomness independent of the history,

$$\mathbb{P}(I(t) = i \mid \mathcal{F}_{t-1}) = \pi_{t,i}. \tag{20}$$

After the initial pull of every arm, let

$$\mathcal{T}_T := \{K+1, \ldots, T\}$$

denote the remaining rounds. The regret decomposes exactly as

$$\begin{aligned}\mathrm{Reg}(T) = &\sum_{j=2}^{K} \Delta_j + \sum_{j=2}^{K} \Delta_j \sum_{t \in \mathcal{T}_T} \mathbf{1}\{I(t) = j,\ \hat{\mu}_1(t) < \mu_1 - \varepsilon\} \\ &+ \sum_{j=2}^{K} \Delta_j \sum_{t \in \mathcal{T}_T} \mathbf{1}\{I(t) = j,\ \hat{\mu}_1(t) \geq \mu_1 - \varepsilon\}.\end{aligned} \tag{21}$$

The first term is the initialization regret. We bound the underestimation and non-underestimation terms in turn.

B.2 Two EI tools

We first record the Gaussian positive-part formula used throughout both regret proofs.

Lemma 1 (Gaussian positive-part identity)**.** *If $X \sim \mathcal{N}(\delta, \sigma^2)$ with $\sigma > 0$, then*

$$\mathbb{E}[X_+] = \sigma\phi(\delta/\sigma) + \delta\Phi(\delta/\sigma). \tag{22}$$

Proof. Write $X = \delta + \sigma Z$ with $Z \sim \mathcal{N}(0,1)$. Direct integration gives

$$\begin{aligned}\mathbb{E}[X_+] &= \int_{-\delta/\sigma}^{\infty} (\delta + \sigma z)\phi(z)\,\mathrm{d}z \\ &= \delta \int_{-\delta/\sigma}^{\infty} \phi(z)\,\mathrm{d}z + \sigma \int_{-\delta/\sigma}^{\infty} z\phi(z)\,\mathrm{d}z \\ &= \delta\Phi(\delta/\sigma) + \sigma\phi(-\delta/\sigma) \\ &= \delta\Phi(\delta/\sigma) + \sigma\phi(\delta/\sigma),\end{aligned}$$

where the third equality uses $\phi'(z) = -z\phi(z)$ and the last uses the symmetry of ϕ. □

For distinct arms $i \neq j$, apply Lemma 1 to $X = \theta_i - \theta_j \sim \mathcal{N}(\delta_{ij}(t), \sigma_{ij}^2(t))$ to obtain Equation (6).

We begin with a pairwise Gaussian estimate that will be used to control the marginal EI. Fix a round and suppress its index. Write $m_i := \hat{\mu}_i(t)$ and $v_i := 1/N_i(t)$, so

$$\theta_i \sim \mathcal{N}(m_i, v_i)$$

independently across different arms. Reusing the definition from Equation (6), abbreviate

$$G_{ij} := G_{ij}(t) = \mathbb{E}[(\theta_i - \theta_j)_+], \qquad i \neq j.$$

We set $G_{ii} := 0$ because repeated virtual selections use the same decision sample rather than independent copies.

Lemma 2 (Pairwise Gaussian product)**.** *For every $i, j \in [K]$,*

$$G_{ij}G_{ji} \leq v_i + v_j.$$

Proof. The case $i = j$ is immediate. For $i \neq j$, let

$$X := \theta_i - \theta_j \sim \mathcal{N}(d, \sigma^2), \qquad \sigma^2 := v_i + v_j.$$

After interchanging i and j if necessary, assume $d \geq 0$, and set

$$a := \mathbb{E}[X_+], \qquad b := \mathbb{E}[(-X)_+].$$

Since $a - b = \mathbb{E}[X] = d$, we have $a = d + b$. If $0 \leq d \leq \sigma$, monotonicity of the positive part gives

$$b \leq \mathbb{E}[(-Z)_+] = \frac{\sigma}{\sqrt{2\pi}}, \qquad Z \sim \mathcal{N}(0, \sigma^2),$$

and hence

$$ab \leq \left(\sigma + \frac{\sigma}{\sqrt{2\pi}}\right) \frac{\sigma}{\sqrt{2\pi}} < \sigma^2.$$

If $d > \sigma$, write $X = d + Z$. The scalar inequality

$$(y - d)_+ \leq \frac{y^2}{4d}$$

and the symmetry of $Z \sim \mathcal{N}(0, \sigma^2)$ give

$$b = \mathbb{E}[(-d - Z)_+] = \mathbb{E}[(Z - d)_+] \leq \frac{\mathbb{E}[Z^2]}{4d} = \frac{\sigma^2}{4d}.$$

Therefore

$$ab \leq \left(d + \frac{\sigma^2}{4d}\right) \frac{\sigma^2}{4d} \leq \frac{5}{16}\sigma^2.$$

Thus $G_{ij}G_{ji} = ab \leq v_i + v_j$ in both cases. □

Let λ_t denote the common marginal EI of the arms satisfying $\pi_{t,i} > 0$, as given by Proposition 1. An arm with $\pi_{t,i} = 0$ has marginal EI at most λ_t.

Lemma 3 (Upper bound on the common marginal EI)**.** *At every decision round,*

$$\lambda_t^2 \leq 2 \sum_{i=1}^{K} \frac{\pi_{t,i}}{N_i(t)}. \tag{23}$$

Proof. Fix t and abbreviate $\pi := \pi_t$. Draw $\boldsymbol{\theta} \sim \Pi_{t-1}$ and, independently, sample $r - 1 = M - 2$ arm indices $A_1, \ldots, A_{r-1}$ according to π. Write

$$W_{r-1} := \max_{\ell \leq r-1} \theta_{A_\ell},$$

with $W_0 := -\infty$ when $M = 2$. Define

$$R_j := \mathbb{E}_{\boldsymbol{\theta}, A_{1:r-1}}[\max\{\theta_j, W_{r-1}\}],$$
$$Q_{ij} := \mathbb{E}_{\boldsymbol{\theta}, A_{1:r-1}}[\max\{\theta_i, \theta_j, W_{r-1}\}],$$

and

$$H_{ij} := H_{ij}(t) = Q_{ij} - R_j.$$

By definition,

$$H_{ij} = \mathbb{E}_{\boldsymbol{\theta}, A_{1:r-1}}[(\theta_i - \max\{\theta_j, W_{r-1}\})_+].$$

Conditioning on the last of the r other virtual arm indices gives

$$\bar{s}_{t,i} = \sum_{j=1}^{K} \pi_j \mathbb{E}_{\boldsymbol{\theta}, A_{1:r-1}}[\max\{\theta_i, \theta_j, W_{r-1}\} - \max\{\theta_j, W_{r-1}\}] = \sum_{j=1}^{K} \pi_j H_{ij}. \tag{24}$$

The matrix Q is symmetric. Put

$$\overline{R} := \sum_j \pi_j R_j.$$

For every i with $\pi_i > 0$, Equation (24) and EI balance give

$$\sum_j \pi_j Q_{ij} = \lambda_t + \overline{R} =: \kappa.$$

Let I, J be independent draws from π and set $Y := Q_{IJ}$. Because $\pi_I > 0$ and $\pi_J > 0$ almost surely and Q is symmetric,

$$\mathbb{E}[Y \mid I] = \mathbb{E}[Y \mid J] = \kappa.$$

Consequently,

$$\begin{aligned}
\sum_{i,j} \pi_i \pi_j H_{ij} H_{ji} &= \mathbb{E}[(Y - R_J)(Y - R_I)] \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y R_I] - \mathbb{E}[Y R_J] + \mathbb{E}[R_I R_J] \\
&= \mathbb{E}[Y^2] - 2\kappa\overline{R} + \overline{R}^2 \\
&= \mathrm{Var}(Y) + \kappa^2 - 2\kappa\overline{R} + \overline{R}^2 \\
&= \mathrm{Var}(Y) + (\kappa - \overline{R})^2 \\
&= \mathrm{Var}(Y) + \lambda_t^2 \geq \lambda_t^2.
\end{aligned}$$

Here $\mathbb{E}[Y R_I] = \mathbb{E}[R_I \mathbb{E}(Y \mid I)] = \kappa\overline{R}$, $\mathbb{E}[Y R_J] = \kappa\overline{R}$, and independence of I and J gives $\mathbb{E}[R_I R_J] = \overline{R}^2$.

Moreover, $0 \leq H_{ij} \leq G_{ij}$ because $\max\{\theta_j, W_{r-1}\} \geq \theta_j$. Lemma 2, with $v_i = 1/N_i(t)$, now gives

$$\begin{aligned}
\lambda_t^2 &\leq \sum_{i,j} \pi_i \pi_j G_{ij} G_{ji} \\
&\leq \sum_{i,j} \pi_i \pi_j \left(\frac{1}{N_i(t)} + \frac{1}{N_j(t)} \right) = 2 \sum_i \frac{\pi_{t,i}}{N_i(t)}.
\end{aligned}$$

□

B.3 Bounding the underestimation term

By Equation (19), whenever $N_1(t) = n$, $\hat{\mu}_1(t) = \hat{\mu}_{1,n}$ and $\hat{\mu}_{1,n} \sim \mathcal{N}(\mu_1, 1/n)$.

For $n \geq 1$ and $x \in \mathbb{R}$, define the threshold expected improvement

$$\mathrm{EI}_\varepsilon(n, x) := \mathbb{E}_{\theta \sim \mathcal{N}(x, 1/n)} \big[(\theta - (\mu_1 - \varepsilon))_+ \big]. \tag{25}$$

Lemma 4 (Gaussian threshold expected-improvement bound)**.** *Let ϕ and Φ denote the standard normal density and cumulative distribution function. For $y \geq 0$, define*

$$h(u) := \phi(u) - u\Phi(-u), \qquad u \geq 0.$$

Then

$$\mathrm{EI}_\varepsilon(n, \mu_1 - \varepsilon - y) = \frac{1}{\sqrt{n}} h(\sqrt{n}\, y), \tag{26}$$

and there is a universal constant $a_1 > 0$ such that

$$\mathrm{EI}_\varepsilon(n, \mu_1 - \varepsilon - y) \geq \frac{a_1}{\sqrt{n}(1 + ny^2)} \exp\left(-\frac{ny^2}{2} \right). \tag{27}$$

Also, for every $x \leq \mu_1 - \varepsilon$,

$$\mathrm{EI}_\varepsilon(n, x) \leq \frac{1}{\sqrt{2\pi n}}. \tag{28}$$

Proof. Write

$$\theta - (\mu_1 - \varepsilon) = \frac{Z - \sqrt{n}\, y}{\sqrt{n}}, \qquad Z \sim \mathcal{N}(0, 1).$$

Applying Lemma 1 to $Z - u \sim \mathcal{N}(-u, 1)$ with $u = \sqrt{n}\, y$ proves Equation (26). Moreover,

$$h(u) = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[(Z-u)_+] = \int_0^\infty v\phi(u+v)\,\mathrm{d}v.$$

Let $\delta_u := (1+u)^{-1}$. For $0 \le v \le \delta_u$, the ratio $\phi(u+v)/\phi(u)$ is at least $\exp(-3/2)$. Indeed,

$$\frac{\phi(u+v)}{\phi(u)} = \exp\left(-uv - \frac{v^2}{2}\right) \ge \exp(-3/2),$$

because $uv \le u/(1+u) \le 1$ and $v^2/2 \le 1/2$. Therefore

$$\begin{aligned} h(u) &\ge \int_0^{\delta_u} v\phi(u+v)\,\mathrm{d}v \\ &\ge \exp(-3/2)\phi(u)\frac{\delta_u^2}{2} \\ &\ge \frac{\exp(-3/2)}{4}\frac{\phi(u)}{1+u^2}. \end{aligned}$$

The last inequality uses $(1+u)^2 \le 2(1+u^2)$, so $\delta_u^2/2 \ge 1/[4(1+u^2)]$. Substituting $u = \sqrt{n}\, y$ proves (27). Finally, the threshold expected improvement is increasing in its Gaussian mean, so for $x \le \mu_1 - \varepsilon$ it is at most its value at $x = \mu_1 - \varepsilon$, namely $1/\sqrt{2\pi n}$. □

For each round, put

$$n_t := N_1(t), \qquad \mathrm{EI}_{\varepsilon,t} := \mathrm{EI}_\varepsilon(n_t, \hat{\mu}_1(t)), \qquad \ell_r := 1 + \log r,$$

and define

$$\alpha_r := \frac{1}{4r}, \qquad \beta_r := \frac{\ell_r}{4r}, \qquad R_{e,t} := \frac{8\ell_r}{\mathrm{EI}_{\varepsilon,t}^2}.$$

The sufficiently sampled, optimistic, and remaining sufficiently sampled sets are

$$\mathcal{S}(t) := \{i \in [K] : N_i(t) \ge \max\{R_{e,t}, n_t + 1\}\}, \tag{29}$$

$$\mathcal{H}(t) := \{i \in \mathcal{S}(t) : i \ne 1,\ \hat{\mu}_i(t) > \mu_1 - \varepsilon\}, \tag{30}$$

$$\mathcal{B}(t) := \mathcal{S}(t) \setminus \mathcal{H}(t). \tag{31}$$

Write

$$\pi_t^{\mathcal{S}} := \sum_{i\in\mathcal{S}(t)} \pi_{t,i}, \qquad \pi_t^{\mathcal{H}} := \sum_{i\in\mathcal{H}(t)} \pi_{t,i},$$

and define the events

$$\mathcal{X}_u(t) := \{\pi_t^{\mathcal{H}} \ge \alpha_r\}, \qquad \mathcal{Y}_u(t) := \{\pi_t^{\mathcal{S}} \ge 1 - \beta_r\}.$$

Thus the underestimation rounds are split as

$$\begin{aligned} &\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\hat{\mu}_1(t) \le \mu_1 - \varepsilon\} \\ &\quad= \underbrace{\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\hat{\mu}_1(t) \le \mu_1 - \varepsilon,\ \mathcal{X}_u(t)\}}_{(\mathrm{L_{u,1}})} + \underbrace{\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\hat{\mu}_1(t) \le \mu_1 - \varepsilon,\ \mathcal{X}_u^c(t)\}}_{(\mathrm{L_{u,2}})}. \end{aligned} \tag{32}$$

B.3.1 Lower bound on the optimal-arm marginal expected improvement

Lemma 5 (Lower bound on $\bar{s}_{t,1}$)**.** *On*

$$\{N_1(t) = n,\ \hat{\mu}_{1,n} \le \mu_1 - \varepsilon\} \cap \mathcal{X}_u^c(t) \cap \mathcal{Y}_u(t),$$

one has

$$\bar{s}_{t,1} \ge \frac{\mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n})}{2\mathrm{e}\sqrt{r}}. \tag{33}$$

Proof. During the block $N_1(t) = n$, write

$$\mathrm{EI}_{\varepsilon,n} := \mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n}), \qquad R_e := \frac{8\ell_r}{\mathrm{EI}_{\varepsilon,n}^2}.$$

By Equation (28),

$$R_e \geq 16\pi\ell_r n > n + 1. \tag{34}$$

Thus arm 1 does not belong to $\mathcal{S}(t)$ and hence does not belong to $\mathcal{B}(t)$.

On $\mathcal{X}_u^c(t) \cap \mathcal{Y}_u(t)$,

$$\sum_{i\in\mathcal{B}(t)} \pi_{t,i} > 1 - \alpha_r - \beta_r = 1 - \frac{2 + \log r}{4r} =: b_r. \tag{35}$$

To lower-bound the marginal EI, consider the event that all r virtual arms are sampled from $\mathcal{B}(t)$. Conditional on their realized indices $a_1, \ldots, a_r \in \mathcal{B}(t)$, define

$$W := \max_{\ell\leq r} \theta_{a_\ell}.$$

The selected arms may repeat. Since they all belong to $\mathcal{B}(t)$, each decision sample θ_{a_ℓ} has mean at most $\mu_1 - \varepsilon$ and variance at most $1/R_e$. Write these means and variances as m_{a_ℓ} and v_{a_ℓ}. For every $\eta > 0$, log-sum-exp, Jensen's inequality, and the Gaussian moment-generating function give

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}\sim\Pi_{t-1}}[(W - (\mu_1 - \varepsilon))_+] &\leq \frac{1}{\eta}\mathbb{E}_{\boldsymbol{\theta}\sim\Pi_{t-1}} \log\left[1 + \sum_{\ell=1}^{r} \exp\{\eta(\theta_{a_\ell} - (\mu_1 - \varepsilon))\}\right] \\
&\leq \frac{1}{\eta}\log \mathbb{E}_{\boldsymbol{\theta}\sim\Pi_{t-1}}\left[1 + \sum_{\ell=1}^{r} \exp\{\eta(\theta_{a_\ell} - (\mu_1 - \varepsilon))\}\right] \\
&= \frac{1}{\eta}\log\left[1 + \sum_{\ell=1}^{r} \exp\left\{\eta(m_{a_\ell} - (\mu_1 - \varepsilon)) + \frac{\eta^2 v_{a_\ell}}{2}\right\}\right] \\
&\leq \frac{1}{\eta}\log\left[1 + r\exp\left(\frac{\eta^2}{2R_e}\right)\right] \\
&\leq \frac{\log(r+1)}{\eta} + \frac{\eta}{2R_e}.
\end{aligned}$$

Here $m_{a_\ell} \leq \mu_1 - \varepsilon$ and $v_{a_\ell} \leq 1/R_e$, and the last step uses $1 + re^c \leq (r+1)e^c$ for $c \geq 0$. To minimize the right-hand side, define

$$f(\eta) := \frac{\log(r+1)}{\eta} + \frac{\eta}{2R_e}, \qquad \eta > 0.$$

Its derivative is

$$f'(\eta) = -\frac{\log(r+1)}{\eta^2} + \frac{1}{2R_e}.$$

Since $f''(\eta) = 2\log(r+1)/\eta^3 > 0$, the unique minimizer is

$$\eta^\star = \sqrt{2R_e \log(r+1)}.$$

Substituting $\eta^\star$ gives

$$\mathbb{E}_{\boldsymbol{\theta}\sim\Pi_{t-1}}[(W - (\mu_1 - \varepsilon))_+] \leq \sqrt{\frac{2\log(r+1)}{R_e}} \leq \frac{\mathrm{EI}_{\varepsilon,n}}{2}. \tag{36}$$

The final inequality follows from $R_e = 8\ell_r/\mathrm{EI}_{\varepsilon,n}^2$ and $\log(r+1) \leq \ell_r$.

For all real x, w,

$$(x - w)_+ \geq (x - (\mu_1 - \varepsilon))_+ - (w - (\mu_1 - \varepsilon))_+. \tag{37}$$

For each fixed index tuple, applying Equation (37) with $x = \theta_1$ and $w = W$, taking expectations term by term, and using Equations (25) and (36) gives

$$\mathbb{E}_{\boldsymbol{\theta}\sim\Pi_{t-1}}[(\theta_1 - W)_+] \geq \frac{\mathrm{EI}_{\varepsilon,n}}{2}.$$

This bound holds for every tuple $a_1, \dots, a_r \in \mathcal{B}(t)$. Let

$$\mathcal{E}_B := \bigcap_{\ell=1}^{r} \{A_\ell \in \mathcal{B}(t)\}.$$

Since the virtual indices are sampled independently according to π_t,

$$\mathbb{P}(\mathcal{E}_B) = \pi_t(\mathcal{B}(t))^r \geq b_r^r.$$

Decomposing the marginal EI over $\mathcal{E}_B$ and its complement gives

$$\begin{aligned}\bar{s}_{t,1} &= \mathbb{E}[\mathbf{1}\{\mathcal{E}_B\}(\theta_1 - W)_+] + \mathbb{E}[\mathbf{1}\{\mathcal{E}_B^c\}(\theta_1 - W)_+] \\ &\geq \mathbb{E}[\mathbf{1}\{\mathcal{E}_B\}(\theta_1 - W)_+] \\ &= \mathbb{P}(\mathcal{E}_B)\mathbb{E}[(\theta_1 - W)_+ \mid \mathcal{E}_B] \\ &\geq \frac{b_r^r \mathrm{EI}_{\varepsilon,n}}{2}.\end{aligned}$$

It remains to bound this action-index probability. Let

$$z_r := \frac{2 + \log r}{4r}.$$

For every $r \geq 1$, $0 \leq z_r \leq 1/2$, and $\log(1 - z) \geq -2z$ on this interval. Hence

$$b_r^r = (1 - z_r)^r \geq \exp(-2r z_r) = \frac{1}{\mathrm{e}\sqrt{r}},$$

which proves Equation (33). □

B.3.2 Number of rounds in a fixed count block

Let

$$\sigma_n := \inf\{t \geq K + 1 : N_1(t) = n\}, \qquad \inf \varnothing := \infty, \tag{38}$$
$$\overline{\sigma}_n := \min\{\sigma_n, T + 1\}, \qquad \mathcal{A}_n := \{\sigma_n \leq T,\ \hat{\mu}_{1,n} \leq \mu_1 - \varepsilon\}.$$

We next bound the conditional expected number of non-optimistic rounds in the count-n block. Define the full count-n non-optimistic block

$$L_n := \sum_{t \in \mathcal{T}_T} \mathbf{1}\{N_1(t) = n,\ \mathcal{X}_u^c(t)\}.$$

Lemma 6 (Fixed-count bound)**.** *One has*

$$\mathbf{1}\{\mathcal{A}_n\}\mathbb{E}[L_n \mid \mathcal{F}_{\overline{\sigma}_n - 1}] \leq \mathbf{1}\{\mathcal{A}_n\} \frac{128 r K \log(T + 1)}{\mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n})^2}. \tag{39}$$

Proof. Work on $\mathcal{A}_n$, and write

$$\mathrm{EI}_{\varepsilon,n} := \mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n}), \qquad R_e := \frac{8\ell_r}{\mathrm{EI}_{\varepsilon,n}^2}.$$

Consider a round counted by L_n. If $\mathcal{Y}_u^c(t)$ occurs, action mass greater than β_r lies outside $\mathcal{S}(t)$. Equation (34) shows that, on this count block, $i \notin \mathcal{S}(t)$ implies $N_i(t) < R_e$. Therefore

$$\sum_i \frac{\pi_{t,i}}{N_i(t)} \geq \frac{\beta_r}{R_e} = \frac{\mathrm{EI}_{\varepsilon,n}^2}{32r}. \tag{40}$$

If $\mathcal{Y}_u(t)$ occurs, Lemma 5 and EI balance imply

$$\lambda_t \geq \bar{s}_{t,1} \geq \frac{\mathrm{EI}_{\varepsilon,n}}{2\mathrm{e}\sqrt{r}}.$$

Lemma 3 then gives

$$\sum_i \frac{\pi_{t,i}}{N_i(t)} \geq \frac{\lambda_t^2}{2} \geq \frac{\mathrm{EI}_{\varepsilon,n}^2}{8\mathrm{e}^2 r} \geq \frac{\mathrm{EI}_{\varepsilon,n}^2}{64r}. \tag{41}$$

Thus every counted round satisfies the common lower bound

$$\sum_i \frac{\pi_{t,i}}{N_i(t)} \geq \frac{\mathrm{EI}_{\varepsilon,n}^2}{64r}. \tag{42}$$

Use the log-count potential

$$\Psi_t := \sum_{i=1}^{K} \log N_i(t).$$

If arm i is pulled at round t, then

$$\Psi_{t+1} - \Psi_t = \log\left(1 + \frac{1}{N_i(t)}\right) \geq \frac{1}{2N_i(t)}.$$

For a deterministic round t, define the following indicator, which is determined before $I(t)$ is drawn:

$$C_{t,n} := \mathbf{1}\{N_1(t) = n,\ \hat{\mu}_1(t) = \hat{\mu}_{1,n} \leq \mu_1 - \varepsilon,\ \mathcal{X}_u^c(t)\}.$$

Using Equation (42),

$$\mathbb{E}[C_{t,n}(\Psi_{t+1} - \Psi_t) \,|\, \mathcal{F}_{t-1}] \geq \frac{C_{t,n}}{2} \sum_i \frac{\pi_{t,i}}{N_i(t)} \geq \frac{\mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n})^2}{128r} C_{t,n}. \tag{43}$$

Fix a deterministic $s \in \mathcal{T}_T$ and define

$$\mathcal{E}_{n,s} := \{\sigma_n = s,\ \hat{\mu}_{1,n} \leq \mu_1 - \varepsilon\} \in \mathcal{F}_{s-1}.$$

Indeed, $\{\sigma_n = s\}$ is determined by the actions through round $s-1$, so σ_n is a stopping time for the shifted filtration $(\mathcal{F}_{t-1})_{t \geq K+1}$. The empirical mean in $\mathcal{E}_{n,s}$ is also determined by those observations. On $\mathcal{E}_{n,s}$, $\mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n})$ is $\mathcal{F}_{s-1}$-measurable, and the definition of $C_{t,n}$ gives

$$\mathbf{1}\{\mathcal{E}_{n,s}\} L_n = \mathbf{1}\{\mathcal{E}_{n,s}\} \sum_{t=s}^{T} C_{t,n}.$$

For every $t \geq s$, multiply Equation (43) by $\mathbf{1}\{\mathcal{E}_{n,s}\}$, condition down to $\mathcal{F}_{s-1}$, and sum over $t = s, \ldots, T$. The tower property gives

$$\begin{aligned}
&\mathbf{1}\{\mathcal{E}_{n,s}\} \frac{\mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n})^2}{128r} \mathbb{E}[L_n \mid \mathcal{F}_{s-1}] \\
&= \mathbf{1}\{\mathcal{E}_{n,s}\} \frac{\mathrm{EI}_\varepsilon(n, \hat{\mu}_{1,n})^2}{128r} \mathbb{E}\left[\sum_{t=s}^{T} C_{t,n} \,\middle|\, \mathcal{F}_{s-1}\right] \\
&\leq \mathbf{1}\{\mathcal{E}_{n,s}\} \mathbb{E}\left[\sum_{t=s}^{T} C_{t,n}(\Psi_{t+1} - \Psi_t) \,\middle|\, \mathcal{F}_{s-1}\right] \\
&\leq \mathbf{1}\{\mathcal{E}_{n,s}\} \mathbb{E}[\Psi_{T+1} - \Psi_s \mid \mathcal{F}_{s-1}] \\
&\leq \mathbf{1}\{\mathcal{E}_{n,s}\} K \log(T+1).
\end{aligned}$$

The events $\{\mathcal{E}_{n,s} : s \in \mathcal{T}_T\}$ are disjoint and their union is $\mathcal{A}_n$. They also show directly that $\mathcal{A}_n \in \mathcal{F}_{\overline{\sigma}_n - 1}$. Decomposing the conditional expectation over the disjoint events $\{E_{n,s} : s \in \mathcal{T}_T\}$ gives,

$$\mathbf{1}\{\mathcal{A}_n\} \mathbb{E}[L_n \mid \mathcal{F}_{\overline{\sigma}_n - 1}] = \sum_{s \in \mathcal{T}_T} \mathbf{1}\{\mathcal{E}_{n,s}\} \mathbb{E}[L_n \mid \mathcal{F}_{s-1}]. \tag{44}$$

Summing the preceding deterministic-s inequalities and applying this identity proves Equation (39). □

B.3.3 First part of the underestimation split

Proposition 2 (Bound for rounds satisfying $\mathcal{X}_u^c(t)$)**.** *There is a universal constant $C > 0$ such that*

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\hat{\mu}_1(t) \le \mu_1 - \varepsilon,\ \mathcal{X}_u^c(t)\}\right] \le C\sqrt{rKT\log(T+1)}\frac{\log(2/\varepsilon)}{\varepsilon^3}. \tag{45}$$

Proof. Set

$$B_{T,M} := 128rK\log(T+1).$$

The left-hand side of Equation (45) is

$$\sum_{n\ge 1}\mathbb{E}[\mathbf{1}\{\mathcal{A}_n\}L_n].$$

Lemma 6, together with $L_n \le T$, gives on $\mathcal{A}_n$

$$\mathbb{E}[L_n \,|\, \mathcal{F}_{\overline{\sigma}_n-1}] \le \min\left\{T, \frac{B_{T,M}}{\mathrm{EI}_\varepsilon(n,\hat{\mu}_{1,n})^2}\right\} \le \frac{\sqrt{B_{T,M}T}}{\mathrm{EI}_\varepsilon(n,\hat{\mu}_{1,n})}. \tag{46}$$

Since $\mathcal{A}_n \in \mathcal{F}_{\overline{\sigma}_n-1}$, the tower property and Equation (46) give

$$\begin{aligned}\mathbb{E}[\mathbf{1}\{\mathcal{A}_n\}L_n] &= \mathbb{E}[\mathbf{1}\{\mathcal{A}_n\}\mathbb{E}[L_n \,|\, \mathcal{F}_{\overline{\sigma}_n-1}]] \\ &\le \sqrt{B_{T,M}T}\,\mathbb{E}\left[\frac{\mathbf{1}\{\mathcal{A}_n\}}{\mathrm{EI}_\varepsilon(n,\hat{\mu}_{1,n})}\right] \\ &\le \sqrt{B_{T,M}T}\,\mathbb{E}\left[\frac{\mathbf{1}\{\hat{\mu}_{1,n} \le \mu_1 - \varepsilon\}}{\mathrm{EI}_\varepsilon(n,\hat{\mu}_{1,n})}\right],\end{aligned}$$

where the last inequality uses

$$\mathcal{A}_n \subseteq \{\hat{\mu}_{1,n} \le \mu_1 - \varepsilon\}.$$

Summing over n gives

$$\sum_{n\ge 1}\mathbb{E}[\mathbf{1}\{\mathcal{A}_n\}L_n] \le \sqrt{B_{T,M}T}\sum_{n\ge 1}\mathbb{E}\left[\frac{\mathbf{1}\{\hat{\mu}_{1,n} \le \mu_1 - \varepsilon\}}{\mathrm{EI}_\varepsilon(n,\hat{\mu}_{1,n})}\right]. \tag{47}$$

Since $\hat{\mu}_{1,n} \sim \mathcal{N}(\mu_1, 1/n)$, substitute

$$\hat{\mu}_{1,n} = \mu_1 - \varepsilon - y, \qquad y \ge 0.$$

The fixed-n expectation in Equation (47) equals

$$\int_0^\infty \frac{f_{\hat{\mu}_{1,n}}(\mu_1 - \varepsilon - y)}{\mathrm{EI}_\varepsilon(n, \mu_1 - \varepsilon - y)}\,\mathrm{d}y.$$

Here

$$\begin{aligned}f_{\hat{\mu}_{1,n}}(\mu_1 - \varepsilon - y) &= \sqrt{\frac{n}{2\pi}}\exp\Big(-\frac{n}{2}(\varepsilon + y)^2\Big) \\ &= \sqrt{\frac{n}{2\pi}}\exp\left(-\frac{n\varepsilon^2}{2} - n\varepsilon y - \frac{ny^2}{2}\right).\end{aligned}$$

Using the preceding density formula and Equation (27), for each $n \ge 1$,

$$\sqrt{B_{T,M}T}\,\mathbb{E}\left[\frac{\mathbf{1}\{\hat{\mu}_{1,n} \le \mu_1 - \varepsilon\}}{\mathrm{EI}_\varepsilon(n,\hat{\mu}_{1,n})}\right] \le C\sqrt{B_{T,M}T}\,n\exp(-n\varepsilon^2/2)\int_0^\infty (1 + ny^2)\exp(-n\varepsilon y)\,\mathrm{d}y.$$

Using the standard Gamma-integral identity

$$\int_0^\infty y^k e^{-ay}\,\mathrm{d}y = \frac{k!}{a^{k+1}}, \qquad a > 0, \quad k \in \{0, 1, 2, \ldots\},$$

with $a = n\varepsilon$ and $k \in \{0, 2\}$, we obtain

$$\int_0^\infty (1 + ny^2)e^{-n\varepsilon y}\,\mathrm{d}y = \int_0^\infty e^{-n\varepsilon y}\,\mathrm{d}y + n\int_0^\infty y^2 e^{-n\varepsilon y}\,\mathrm{d}y = \frac{1}{n\varepsilon} + \frac{2}{n^2\varepsilon^3}. \tag{48}$$

For each $n \geq 1$, the term indexed by n on the right-hand side of Equation (47) is bounded by

$$C\sqrt{B_{T,M}T}\,\exp(-n\varepsilon^2/2)\left(\frac{1}{\varepsilon} + \frac{2}{n\varepsilon^3}\right).$$

Therefore, summing this bound over n gives

$$\sum_{n\geq 1}\mathbb{E}[\mathbf{1}\{\mathcal{A}_n\}L_n] \leq C\sqrt{B_{T,M}T}\left[\frac{1}{\varepsilon}\sum_{n\geq 1}\exp(-n\varepsilon^2/2) + \frac{2}{\varepsilon^3}\sum_{n\geq 1}\frac{\exp(-n\varepsilon^2/2)}{n}\right].$$

For $\varepsilon \leq 1$, the two series satisfy

$$\sum_{n\geq 1}\exp(-n\varepsilon^2/2) = \frac{1}{\exp(\varepsilon^2/2) - 1} \leq \frac{2}{\varepsilon^2},$$

and

$$\sum_{n\geq 1}\frac{\exp(-n\varepsilon^2/2)}{n} = -\log\bigl(1 - \exp(-\varepsilon^2/2)\bigr) \leq C\log(2/\varepsilon).$$

Substituting these bounds into the preceding inequality yields

$$\begin{aligned}\sum_{n\geq 1}\mathbb{E}[\mathbf{1}\{\mathcal{A}_n\}L_n] &\leq C\sqrt{B_{T,M}T}\left[\frac{1}{\varepsilon^3} + \frac{\log(2/\varepsilon)}{\varepsilon^3}\right] \\ &\leq C\sqrt{B_{T,M}T}\frac{\log(2/\varepsilon)}{\varepsilon^3}.\end{aligned}$$

Finally, substituting $B_{T,M} = 128rK\log(T+1)$ and absorbing the numerical factor into C gives

$$\sum_{n\geq 1}\mathbb{E}[\mathbf{1}\{\mathcal{A}_n\}L_n] \leq C\sqrt{rKT\log(T+1)}\frac{\log(2/\varepsilon)}{\varepsilon^3},$$

which is Equation (45). □

B.3.4 Second part of the underestimation split

Proposition 3 (Bound for rounds satisfying $\mathcal{X}_u(t)$)**.** *One has*

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_T}\mathbf{1}\{\hat{\mu}_1(t) \leq \mu_1 - \varepsilon,\ \mathcal{X}_u(t)\}\right] \leq 4r\sum_{i=2}^{K}\frac{1}{\exp((\Delta_i - \varepsilon)^2/2) - 1}. \tag{49}$$

Proof. On $\mathcal{X}_u(t)$, action mass at least $\alpha_r = 1/(4r)$ lies on $\mathcal{H}(t)$, and hence

$$\mathbf{1}\{\mathcal{X}_u(t)\} \leq \frac{1}{\alpha_r}\sum_{i=2}^{K}\pi_{t,i}\mathbf{1}\{i \in \mathcal{H}(t)\}. \tag{50}$$

All quantities in Equation (50), as well as the indicator $\mathbf{1}\{\widehat{\mu}_1(t) \leq \mu_1 - \varepsilon\}$, are determined before $I(t)$ is drawn. Multiplying Equation (50) by this indicator, summing over t, and using Equation (20) together with the tower property yields

$$\mathbb{E}\sum_{t\in\mathcal{T}_T}\mathbf{1}\{\hat{\mu}_1(t) \leq \mu_1 - \varepsilon,\ \mathcal{X}_u(t)\} \leq 4r\sum_{i=2}^{K}\mathbb{E}\sum_{t\in\mathcal{T}_T}\mathbf{1}\{I(t) = i,\ \hat{\mu}_i(t) > \mu_1 - \varepsilon\}. \tag{51}$$

For each fixed $i \neq 1$ and $m \geq 1$, arm i can be pulled at most once while its pre-count equals m. On that round, $\hat{\mu}_i(t) = \hat{\mu}_{i,m}$, so the pull is included in the preceding sum only if

$$\{\hat{\mu}_{i,m} > \mu_1 - \varepsilon\}.$$

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

$$\mathbb{P}(\hat{\mu}_{i,m} > \mu_1 - \varepsilon) \le \exp\left(-\frac{m(\Delta_i - \varepsilon)^2}{2}\right).$$

Therefore, for each $i \neq 1$,

$$\mathbb{E}\sum_{t\in\mathcal{T}_T} \mathbf{1}\{I(t) = i,\ \hat{\mu}_i(t) > \mu_1 - \varepsilon\} \le \sum_{m\ge 1} \exp\left(-\frac{m(\Delta_i - \varepsilon)^2}{2}\right) = \frac{1}{\exp((\Delta_i - \varepsilon)^2/2) - 1}.$$

Substituting this bound into the preceding inequality and summing over $i = 2, \dots, K$ proves Equation (49). □

Combining the preceding two propositions bounds the total number of rounds on which the optimal empirical mean is underestimated.

Corollary 1 (Total number of underestimation rounds)**.**

$$\mathbb{E}\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\hat{\mu}_1(t) \le \mu_1 - \varepsilon\} \le C\sqrt{rKT\log(T+1)}\frac{\log(2/\varepsilon)}{\varepsilon^3} + 4r\sum_{i=2}^{K} \frac{1}{\exp((\Delta_i - \varepsilon)^2/2) - 1}. \tag{52}$$

Proof. Split the underestimation event according to $\mathcal{X}_u(t)$ and $\mathcal{X}_u^c(t)$. Applying Propositions 2 and 3 to the two resulting terms and adding their bounds gives the result. □

B.4 Bounding the non-underestimation term

Fix a suboptimal arm $j \ge 2$. Define the following pre-decision event

$$\mathcal{X}_s(t) := \{\hat{\mu}_1(t) \ge \mu_1 - \varepsilon,\ \hat{\mu}_j(t) \le \mu_j + \varepsilon\}. \tag{53}$$

On this event,

$$d_t := \hat{\mu}_1(t) - \hat{\mu}_j(t) \ge \Delta_j - 2\varepsilon \ge \frac{3}{5}\Delta_j > 0, \tag{54}$$

where the second inequality uses $\varepsilon \le \Delta_{\min}/5 \le \Delta_j/5$.

B.4.1 A CDF representation of marginal EI

For a fixed round t and posterior realization $\boldsymbol{\theta}$, define

$$F_{\pi_t}(u; \boldsymbol{\theta}) := \sum_{k=1}^{K} \pi_{t,k}\mathbf{1}\{\theta_k \le u\}.$$

Conditional on $\boldsymbol{\theta}$, the maximum $W_r = W_{M-1}$ of the r other virtual actions satisfies

$$\mathbb{P}(W_r \le u \mid \boldsymbol{\theta}) = F_{\pi_t}(u; \boldsymbol{\theta})^r.$$

Lemma 7 (Marginal-EI CDF identity)**.** *For every arm i,*

$$\bar{s}_{t,i} = \mathbb{E}_{\boldsymbol{\theta}\sim\Pi_{t-1}} \int_{-\infty}^{\theta_i} F_{\pi_t}(u; \boldsymbol{\theta})^r \,\mathrm{d}u. \tag{55}$$

Consequently,

$$\bar{s}_{t,1} - \bar{s}_{t,j} = \mathbb{E}_{\boldsymbol{\theta}\sim\Pi_{t-1}} \int_{\theta_j}^{\theta_1} F_{\pi_t}(u; \boldsymbol{\theta})^r \,\mathrm{d}u. \tag{56}$$

Proof. Let

$$W_r := \max_{\ell \le r} \theta_{A_\ell}, \qquad A_1, \dots, A_r \overset{\text{i.i.d.}}{\sim} \pi_t.$$

Conditional on $\boldsymbol{\theta}$, the event $\{W_r \le u\}$ occurs if and only if every sampled arm satisfies $\theta_{A_\ell} \le u$. Hence, by independence of the virtual action indices,

$$\mathbb{P}(W_r \le u \mid \boldsymbol{\theta}) = \prod_{\ell=1}^{r} \mathbb{P}(\theta_{A_\ell} \le u \mid \boldsymbol{\theta}) = F_{\pi_t}(u; \boldsymbol{\theta})^r.$$

For all real x, w,

$$(x - w)_+ = \int_{-\infty}^{x} \mathbf{1}\{w \le u\} \, \mathrm{d}u.$$

Applying this identity with $x = \theta_i$ and $w = W_r$, and then using Tonelli's theorem to interchange the nonnegative integral and expectation, gives

$$\begin{aligned} \bar{s}_{t,i} &= \mathbb{E}_{\boldsymbol{\theta}, A_{1:r}}[(\theta_i - W_r)_+] \\ &= \mathbb{E}_{\boldsymbol{\theta}} \int_{-\infty}^{\theta_i} \mathbb{P}(W_r \le u \mid \boldsymbol{\theta}) \, \mathrm{d}u \\ &= \mathbb{E}_{\boldsymbol{\theta} \sim \Pi_{t-1}} \int_{-\infty}^{\theta_i} F_{\pi_t}(u; \boldsymbol{\theta})^r \, \mathrm{d}u, \end{aligned}$$

which proves Equation (55).

Applying this identity to arms 1 and j and subtracting yields

$$\begin{aligned} \bar{s}_{t,1} - \bar{s}_{t,j} &= \mathbb{E}_{\boldsymbol{\theta} \sim \Pi_{t-1}} \left[\int_{-\infty}^{\theta_1} F_{\pi_t}(u; \boldsymbol{\theta})^r \, \mathrm{d}u - \int_{-\infty}^{\theta_j} F_{\pi_t}(u; \boldsymbol{\theta})^r \, \mathrm{d}u \right] \\ &= \mathbb{E}_{\boldsymbol{\theta} \sim \Pi_{t-1}} \int_{\theta_j}^{\theta_1} F_{\pi_t}(u; \boldsymbol{\theta})^r \, \mathrm{d}u, \end{aligned}$$

which proves Equation (56). □

B.4.2 THE PROBABILITY OF A SATURATED ARM

For this fixed arm j, define

$$A_{j,M} := \left(\frac{4}{\sqrt{\pi}\, \Delta_j} \right)^{1/r}, \qquad f_t^\star := \frac{d_t^2}{2(1/N_1(t) + 1/N_j(t))}. \tag{57}$$

Lemma 8 (Pairwise saturation inequality)**.** *If $\pi_{t,j} > 0$, then $\pi_{t,j} < 1$ and*

$$\left(\frac{\pi_{t,j}}{1 - \pi_{t,j}} \right)^r \le \frac{G_{j1}(t)}{G_{1j}(t)}. \tag{58}$$

On $\mathcal{X}_s(t)$, for every j,

$$\pi_{t,j} \le A_{j,M} \exp\left(-\frac{f_t^\star}{r} \right). \tag{59}$$

Proof. Suppress the round index t. If $\theta_1 > \theta_j$, then for every $u \in [\theta_j, \theta_1)$, arm j contributes its probability mass π_j to $F_\pi(u; \boldsymbol{\theta})$. Hence

$$\int_{\theta_j}^{\theta_1} F_\pi(u; \boldsymbol{\theta})^r \, \mathrm{d}u \ge \pi_j^r (\theta_1 - \theta_j).$$

If $\theta_j > \theta_1$, then for every $u \in [\theta_1, \theta_j)$, arm j does not contribute to $F_\pi(u; \boldsymbol{\theta})$, so $F_\pi(u; \boldsymbol{\theta}) \le 1 - \pi_j$. Since the integral is taken in the reverse direction,

$$\int_{\theta_j}^{\theta_1} F_\pi(u; \boldsymbol{\theta})^r \, \mathrm{d}u \ge -(1 - \pi_j)^r (\theta_j - \theta_1).$$

Combining the two cases gives the pointwise bound

$$\int_{\theta_j}^{\theta_1} F_\pi(u;\boldsymbol{\theta})^r \,\mathrm{d}u \geq \pi_j^r(\theta_1-\theta_j)_+ - (1-\pi_j)^r(\theta_j-\theta_1)_+.$$

Taking posterior expectations and using the definitions of G_{1j} and G_{j1} in Equation (56) yields

$$\bar{s}_{t,1} - \bar{s}_{t,j} \geq \pi_j^r G_{1j} - (1-\pi_j)^r G_{j1}.$$

If $\pi_{t,j} > 0$, Proposition 1 gives $\bar{s}_{t,j} \geq \bar{s}_{t,1}$, and hence

$$\pi_j^r G_{1j}(t) \leq (1-\pi_j)^r G_{j1}(t).$$

The Gaussian laws make both pairwise EIs strictly positive, so this inequality rules out $\pi_j = 1$. Division proves Equation (58).

On $\mathcal{X}_s(t)$, Jensen's inequality for the convex positive part and Equation (54) give

$$G_{1j}(t) \geq d_t \geq \frac{3}{5}\Delta_j \geq \frac{\Delta_j}{4}. \tag{60}$$

Put

$$q_t := \frac{1}{N_1(t)} + \frac{1}{N_j(t)}.$$

Since $\theta_j - \theta_1 \sim \mathcal{N}(-d_t, q_t)$, Lemma 1 gives

$$\begin{aligned} G_{j1}(t) &= \sqrt{q_t}\,\phi\!\left(\frac{d_t}{\sqrt{q_t}}\right) - d_t\,\Phi\!\left(-\frac{d_t}{\sqrt{q_t}}\right) \\ &\leq \sqrt{q_t}\,\phi\!\left(\frac{d_t}{\sqrt{q_t}}\right) \\ &= \frac{\sqrt{q_t}}{\sqrt{2\pi}}\exp(-f_t^\star) \leq \frac{1}{\sqrt{\pi}}\exp(-f_t^\star), \end{aligned} \tag{61}$$

where the final inequality uses $q_t \leq 2$. Combining Equations (58), (60), and (61), then taking an r-th root, yields

$$\pi_{t,j} \leq \frac{\pi_{t,j}}{1-\pi_{t,j}} \leq \left(\frac{4}{\sqrt{\pi}\,\Delta_j}\right)^{1/r} \exp\!\left(-\frac{f_t^\star}{r}\right).$$

If $\pi_{t,j} = 0$, the same conclusion is immediate. □

B.4.3 Bounding suboptimal-arm pulls via the best arm

Lemma 9 (Best-arm probability dominance)**.** *Suppose*

$$\hat{\mu}_1(t) > \hat{\mu}_j(t), \qquad N_1(t) \leq N_j(t).$$

Then

$$\pi_{t,j} \leq \pi_{t,1}. \tag{62}$$

Proof. Suppress the round index t and suppose, for contradiction, that $\pi_j > \pi_1$. Conditional on a fixed realization of all arms other than 1 and j, define

$$B(u) := \sum_{k\notin\{1,j\}} \pi_k \mathbf{1}\{\theta_k \leq u\}, \qquad H_B(x) := \int_0^x (B(u)+\pi_1)^r \,\mathrm{d}u.$$

Since $B(u)$ is nonnegative and nondecreasing, H_B is increasing and convex.

For u between θ_1 and θ_j, one has $F_\pi(u;\boldsymbol{\theta}) = B(u) + \pi_j$ when $\theta_1 > \theta_j$, and $F_\pi(u;\boldsymbol{\theta}) = B(u) + \pi_1$ when $\theta_j > \theta_1$. Moreover,

$$H_B(\theta_1) - H_B(\theta_j) = \int_{\theta_j}^{\theta_1} (B(u)+\pi_1)^r \,\mathrm{d}u.$$

Therefore, Equation (56) gives

$$\bar{s}_{t,1} - \bar{s}_{t,j} = \mathbb{E}[H_B(\theta_1) - H_B(\theta_j)] + \mathbb{E}\left[\mathbf{1}\{\theta_1 > \theta_j\} \int_{\theta_j}^{\theta_1} \left\{(B(u)+\pi_j)^r - (B(u)+\pi_1)^r\right\} \mathrm{d}u\right]. \tag{63}$$

Let $\delta := \hat{\mu}_1 - \hat{\mu}_j > 0$. Since $1/N_1 \geq 1/N_j$, we may write

$$\theta_1 \overset{d}{=} \theta'_j + \delta + Z,$$

where θ'_j has the same distribution as θ_j, and Z is a centered Gaussian with variance $1/N_1 - 1/N_j$. On the auxiliary probability space, take θ'_j and Z independent of each other and of the already fixed realization of all other arms. For every fixed function B and every fixed value x of θ'_j, conditional Jensen and the monotonicity of H_B give

$$\mathbb{E}_Z[H_B(x+\delta+Z)] \geq H_B(x+\delta+\mathbb{E}Z) = H_B(x+\delta) \geq H_B(x).$$

Integrating this inequality over θ'_j yields

$$\mathbb{E}[H_B(\theta_1)] \geq \mathbb{E}[H_B(\theta'_j+\delta)] \geq \mathbb{E}[H_B(\theta'_j)] = \mathbb{E}[H_B(\theta_j)].$$

Hence the first term in Equation (63) is nonnegative.

On $\{\theta_1 > \theta_j\}$, the second integrand is strictly positive because $\pi_j > \pi_1$ and $r \geq 1$. Since $\mathbb{P}(\theta_1 > \theta_j) > 0$, the second term is strictly positive. Thus

$$\bar{s}_{t,1} > \bar{s}_{t,j}.$$

However, $\pi_j > 0$, so EI balance requires $\bar{s}_{t,j} \geq \bar{s}_{t,1}$, a contradiction. Therefore $\pi_j \leq \pi_1$. □

B.4.4 Counting saturated pulls

We first isolate overestimation of the suboptimal arm.

Lemma 10 (Arm j overestimation bound)**.**

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_T} \mathbf{1}\{I(t)=j,\ \hat{\mu}_1(t) \geq \mu_1 - \varepsilon,\ \hat{\mu}_j(t) > \mu_j + \varepsilon\}\right] \leq \frac{2}{\varepsilon^2}. \tag{64}$$

Proof. Drop the condition on arm 1. For each $m \geq 1$, arm j can be pulled at most once while its pre-count equals m. On that round, $\hat{\mu}_j(t) = \hat{\mu}_{j,m}$, so the pull is counted only if $\hat{\mu}_{j,m} > \mu_j + \varepsilon$. Therefore the expectation is at most

$$\sum_{m\geq 1} \mathbb{P}(\hat{\mu}_{j,m} > \mu_j + \varepsilon) \leq \sum_{m\geq 1} \exp(-m\varepsilon^2/2) \leq \frac{2}{\varepsilon^2}.$$

□

For the saturation argument, define

$$L_{j,M} := 1 + 2TA_{j,M}, \qquad T_{j,M} := \left\lceil \max\left\{1, \frac{r}{D_j^{\varepsilon}} \log L_{j,M}\right\}\right\rceil. \tag{65}$$

Also define

$$\mathcal{Y}_s(t) := \{N_j(t) \geq 3T_{j,M}\}, \qquad \mathcal{Z}_s(t) := \{N_1(t) \geq 3T_{j,M}\}.$$

Lemma 11 (Saturation count)**.** *For every suboptimal arm j,*

$$\mathbb{E}\left[\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\hat{\mu}_1(t) \geq \mu_1 - \varepsilon,\ I(t)=j\}\right] \leq C\left[\frac{1}{\varepsilon^2} + \frac{r}{D_j^{\varepsilon}} \log L_{j,M}\right] \tag{66}$$

for a universal constant $C > 0$.

Proof. Lemma 10 removes the rounds on which arm j is overestimated at a cost at most $2/\varepsilon^2$. It remains to count pulls on $\mathcal{X}_s(t)$.

Before $N_j(t)$ reaches $3T_{j,M}$, arm j can be drawn at most $3T_{j,M}$ times. The remaining pulls split as

$$\sum_{t\in\mathcal{T}_T} \mathbf{1}\{I(t)=j,\ \mathcal{X}_s(t),\ \mathcal{Y}_s(t)\}$$
$$= \underbrace{\sum_{t\in\mathcal{T}_T} \mathbf{1}\{I(t)=j,\ \mathcal{X}_s(t),\ \mathcal{Y}_s(t),\ \mathcal{Z}_s(t)\}}_{(\mathrm{L}_{\mathrm{s},1})} + \underbrace{\sum_{t\in\mathcal{T}_T} \mathbf{1}\{I(t)=j,\ \mathcal{X}_s(t),\ \mathcal{Y}_s(t),\ \mathcal{Z}_s^c(t)\}}_{(\mathrm{L}_{\mathrm{s},2})}.$$

For the rounds counted by $(\mathrm{L}_{\mathrm{s},2})$, $N_1(t) < N_j(t)$ and $d_t > 0$. Lemma 9 therefore gives $\pi_{t,j} \le \pi_{t,1}$. Since $\mathcal{X}_s(t)$, $\mathcal{Y}_s(t)$, and $\mathcal{Z}_s^c(t)$ are determined by $\mathcal{F}_{t-1}$, Equation (20) gives

$$\begin{aligned}\mathbb{E}[(\mathrm{L}_{\mathrm{s},2})] &\le \mathbb{E}\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\mathcal{X}_s(t),\mathcal{Y}_s(t),\mathcal{Z}_s^c(t)\}\pi_{t,1}\\ &= \mathbb{E}\sum_{t\in\mathcal{T}_T} \mathbf{1}\{\mathcal{X}_s(t),\mathcal{Y}_s(t),\mathcal{Z}_s^c(t),\ I(t)=1\}\\ &\le 3T_{j,M}+1,\end{aligned}$$

because arm 1 can be pulled only that many times before its count reaches $3T_{j,M}$.

For rounds counted by $(\mathrm{L}_{\mathrm{s},1})$, both counts are at least $3T_{j,M}$, and hence

$$\frac{1}{1/N_1(t)+1/N_j(t)} \ge \frac{3}{2}T_{j,M}.$$

Equations (54) and (65) give

$$\begin{aligned}\frac{f_t^\star}{r} &= \frac{d_t^2}{2r}\frac{1}{1/N_1(t)+1/N_j(t)}\\ &\ge \frac{3T_{j,M}}{4r}(\Delta_j-2\varepsilon)^2\\ &\ge \frac{3}{4}\frac{(\Delta_j-2\varepsilon)^2}{D_j^\varepsilon}\log L_{j,M}\\ &= q_j\log L_{j,M},\\ q_j &:= \frac{3}{2}\left(\frac{\Delta_j-2\varepsilon}{\Delta_j-\frac{3}{2}\varepsilon}\right)^2.\end{aligned}$$

Since $\varepsilon \le \Delta_j/5$,

$$q_j \ge \frac{3}{2}\left(\frac{3/5}{7/10}\right)^2 = \frac{54}{49} > 1.$$

Lemma 8 now yields

$$\pi_{t,j} \le A_{j,M}L_{j,M}^{-q_j} \le \frac{1}{2T},$$

where the last inequality uses $L_{j,M} \ge 2TA_{j,M}$ and $q_j > 1$. Therefore

$$\mathbb{E}[(\mathrm{L}_{\mathrm{s},1})] \le \sum_{t\in\mathcal{T}_T}\frac{1}{2T} \le \frac{1}{2}.$$

Combining all four contributions explicitly gives

$$\begin{aligned}&\mathbb{E}\left[\sum_{t\in\mathcal{T}_T}\mathbf{1}\{\hat{\mu}_1(t)\ge\mu_1-\varepsilon,\ I(t)=j\}\right]\\ &\le \frac{2}{\varepsilon^2}+3T_{j,M}+(3T_{j,M}+1)+\frac{1}{2}\\ &\le C\left[\frac{1}{\varepsilon^2}+\frac{r}{D_j^\varepsilon}\log L_{j,M}\right].\end{aligned}$$

The last inequality uses $\varepsilon \le 1$ and $T_{j,M} \le 2 + r\log(L_{j,M})/D_j^\varepsilon$. This proves Equation (66). □

B.5 Completion of the proof of Theorem 1

Proof of Theorem 1. Take expectations in the regret decomposition (21). We bound its underestimation and non-underestimation terms separately.

Since at most one suboptimal arm is pulled at each round and $\Delta_j \le \Delta_{\max}$, Corollary 1 gives

$$\begin{aligned}
&\mathbb{E}\sum_{j=2}^{K}\Delta_j \sum_{t\in\mathcal{T}_T}\mathbf{1}\{I(t)=j,\ \hat{\mu}_1(t)<\mu_1-\varepsilon\}\\
&\quad\le \Delta_{\max}\mathbb{E}\sum_{t\in\mathcal{T}_T}\mathbf{1}\{\hat{\mu}_1(t)\le\mu_1-\varepsilon\}\\
&\quad\le C\Delta_{\max}\sqrt{rKT\log(T+1)}\frac{\log(2/\varepsilon)}{\varepsilon^3}+4r\Delta_{\max}\sum_{i=2}^{K}\frac{1}{\exp((\Delta_i-\varepsilon)^2/2)-1}.
\end{aligned}$$

Applying Lemma 11 to each suboptimal arm j gives

$$\begin{aligned}
&\mathbb{E}\sum_{j=2}^{K}\Delta_j \sum_{t\in\mathcal{T}_T}\mathbf{1}\{I(t)=j,\ \hat{\mu}_1(t)\ge\mu_1-\varepsilon\}\\
&\quad\le C\sum_{j=2}^{K}\Delta_j\left[\frac{1}{\varepsilon^2}+\frac{r}{D_j^{\varepsilon}}\log L_{j,M}\right].
\end{aligned}$$

Finally, the initialization regret is $\sum_{j=2}^{K}\Delta_j$. Adding the three bounds and substituting

$$L_{j,M}=1+2T\left(\frac{4}{\sqrt{\pi}\,\Delta_j}\right)^{1/r}$$

gives Equation (7), completing the proof. □

C Proof of Theorem 2

We retain the notation and the pre-decision convention of Appendix B. In particular, $r=M-1$, $N_i(t)$ and $\hat{\mu}_i(t)$ are evaluated before the action at round t, and $\hat{\mu}_{i,m}$ is the average of the first m rewards of arm i. We also use the quantities $G_{ij}(t)$, $H_{ij}(t)$, $\bar{s}_{t,i}$, and λ_t defined in Section 2 and Appendix B.

All constants in this appendix may depend on the fixed mean vector, K, M, and a displayed positive gap. They never depend on the round, the horizon, or the current sample counts. A constant denoted by C may have a different value in different displays.

C.1 Reward sequences and simultaneous deviation bounds

The reward sequences $X_{i,m}$ and their averages $\hat{\mu}_{i,m}$ were defined in Equation (19). Fix $\beta>2$ and define the two nonnegative random variables

$$B_{\mathrm{dev}}:=\max_{i\in[K]}\sup_{m\ge1}\left\{m(\hat{\mu}_{i,m}-\mu_i)^2-\beta\log(em)\right\}_+, \tag{67}$$

$$B_{\mathrm{dev},-1}:=\max_{j>1}\sup_{m\ge1}\left\{m(\hat{\mu}_{j,m}-\mu_j)^2-\beta\log(em)\right\}_+. \tag{68}$$

The definition gives, simultaneously for every arm and every count,

$$|\hat{\mu}_{i,m}-\mu_i|\le\sqrt{\frac{B_{\mathrm{dev}}+\beta\log(em)}{m}}, \tag{69}$$

$$|\hat{\mu}_{j,m}-\mu_j|\le\sqrt{\frac{B_{\mathrm{dev},-1}+\beta\log(em)}{m}},\qquad j>1. \tag{70}$$

For $x \geq 0$, a union bound and the Gaussian tail inequality give

$$\mathbb{P}(B_{\mathrm{dev}} > x) \leq \sum_{i=1}^{K} \sum_{m=1}^{\infty} \mathbb{P}\left(|\hat{\mu}_{i,m} - \mu_i| > \sqrt{\frac{x + \beta \log(em)}{m}} \right) \tag{71}$$

$$\leq 2K \sum_{m=1}^{\infty} \exp\left(-\frac{x + \beta \log(em)}{2} \right) \tag{72}$$

$$= 2Ke^{-x/2} \sum_{m=1}^{\infty} (em)^{-\beta/2}, \tag{73}$$

which is exponentially small in x. The series is finite because $\beta > 2$. Therefore B_{dev} and $B_{\mathrm{dev},-1}$ are finite almost surely, have moments of every polynomial order, and satisfy

$$\mathbb{E}[e^{aB_{\mathrm{dev}}}] < \infty, \qquad \mathbb{E}[e^{aB_{\mathrm{dev},-1}}] < \infty, \qquad 0 < a < \frac{1}{2}. \tag{74}$$

Moreover, $B_{\mathrm{dev},-1}$ is independent of the entire reward sequence of arm 1.

For the proof only, reveal all reward sequences at time zero and define

$$\widetilde{\mathcal{F}}_t := \sigma(\mathcal{F}_t, X_{i,m} : i \in [K],\ m \geq 1)\,. \tag{75}$$

This does not change the algorithm. Its fresh random variable for drawing $I(t)$ is independent of the reward sequences, so

$$\mathbb{P}(I(t) = i \mid \widetilde{\mathcal{F}}_{t-1}) = \pi_{t,i}. \tag{76}$$

let $A_{1:M} \sim \pi$ denote M i.i.d. *virtual arm* indices drawn from π. Likewise, the Gaussian variables and the virtual arm indices used in the definition of $\bar{s}_{t,i}$ can be generated from fresh random variables. Thus, conditional on $\widetilde{\mathcal{F}}_{t-1}$, the $r = M - 1$ virtual arm indices are independent draws from π_t.

Recall the count times σ_n from Equation (38). Under the pre-decision convention, $\sigma_1 = K + 1$, and the rounds

$$t = \sigma_n, \ldots, \sigma_{n+1} - 1 \tag{77}$$

are exactly the rounds on which $N_1(t) = n$ and $\hat{\mu}_1(t) = \hat{\mu}_{1,n}$. We will prove below that every σ_n is finite almost surely. For conditioning at the start of this interval, write

$$\mathcal{G}_n := \widetilde{\mathcal{F}}_{\sigma_n - 1}. \tag{78}$$

Thus, $\mathcal{G}_n$ is the information available immediately before the random round σ_n. The interval estimates below are used only under the induction hypothesis $\sigma_n < \infty$ almost surely; Proposition 4 starts from $\sigma_1 = K + 1$ and verifies this hypothesis successively.

C.2 Pairwise expected improvement between suboptimal arms

The only additional assumption used in this subsection is that distinct suboptimal arms have different means, which is an additional assumption in Theorem 2. Define

$$\gamma_- := \begin{cases} \min\limits_{\substack{i,j>1 \\ i \neq j}} |\mu_i - \mu_j|, & K \geq 3, \\ 1, & K = 2, \end{cases} \tag{79}$$

which is larger than 0. For $K = 2$, every assertion below involving two distinct suboptimal arms is empty.

Lemma 12 (Pairwise EI between well-sampled suboptimal arms)**.** *There are constants $A_\gamma, C_\gamma > 0$ such that, for every round t and all distinct $i, j > 1$, with $m_{ij}(t) := \min\{N_i(t), N_j(t)\}$,*

$$\min\{G_{ij}(t), G_{ji}(t)\} \leq A_\gamma \exp(C_\gamma B_{\mathrm{dev},-1}) \exp\left(-\frac{\gamma_-^2 m_{ij}(t)}{16} \right). \tag{80}$$

Proof. Fix t and distinct $i, j > 1$. Put

$$N_i := N_i(t), \qquad N_j := N_j(t), \qquad m := \min\{N_i, N_j\},$$

and define

$$v_{ij} := \frac{1}{N_i} + \frac{1}{N_j}, \qquad w_{ij} := \frac{1}{v_{ij}} = \frac{N_i N_j}{N_i + N_j}. \tag{81}$$

Then

$$\frac{m}{2} \le w_{ij} \le m. \tag{82}$$

Write $e_k := \hat{\mu}_k(t) - \mu_k$. Suppose first that $N_i \ge N_j = m$. Equation (70) and $w_{ij}/N_i \le m/N_i$ give

$$w_{ij} e_i^2 \le \frac{m}{N_i}\{B_{\mathrm{dev},-1} + \beta \log(eN_i)\} \tag{83}$$

$$\le B_{\mathrm{dev},-1} + \beta \log(em) + \beta \frac{m}{N_i} \log(N_i/m) \tag{84}$$

$$\le B_{\mathrm{dev},-1} + \beta \log(em) + \frac{\beta}{e}, \tag{85}$$

because $\sup_{x \ge 1} (\log x)/x = 1/e$. Also,

$$w_{ij} e_j^2 \le B_{\mathrm{dev},-1} + \beta \log(em). \tag{86}$$

The same two bounds hold after interchanging i and j. Hence

$$w_{ij}(e_i - e_j)^2 \le 2 w_{ij} e_i^2 + 2 w_{ij} e_j^2 \tag{87}$$

$$\le C\{1 + B_{\mathrm{dev},-1} + \log(em)\}. \tag{88}$$

Let $\delta := \mu_i - \mu_j$. Since $(a+b)^2 \ge a^2/2 - b^2$ for all real a, b, Equations (82) and (88) imply

$$w_{ij}(\hat{\mu}_i(t) - \hat{\mu}_j(t))^2 \ge \frac{w_{ij}\delta^2}{2} - w_{ij}(e_i - e_j)^2 \tag{89}$$

$$\ge \frac{m\gamma_-^2}{4} - C\{1 + B_{\mathrm{dev},-1} + \log(em)\}. \tag{90}$$

Set

$$x := \sqrt{w_{ij}}\, |\hat{\mu}_i(t) - \hat{\mu}_j(t)|.$$

The posterior difference has variance $v_{ij} = w_{ij}^{-1}$. Lemma 1 therefore gives

$$\min\{G_{ij}(t), G_{ji}(t)\} = \sqrt{v_{ij}}\{\phi(x) - x\Phi(-x)\} \le \sqrt{v_{ij}}\,\phi(x). \tag{91}$$

Since $v_{ij} \le 2$ and $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$, Equation (90) yields

$$\min\{G_{ij}(t), G_{ji}(t)\} \le C \exp(C B_{\mathrm{dev},-1}) m^C \exp\left(-\frac{\gamma_-^2 m}{8}\right). \tag{92}$$

Finally, for $a := \gamma_-^2/16$,

$$\sup_{m \ge 1} m^C e^{-am} \le \max\left\{1, \left(\frac{C}{ae}\right)^C\right\}. \tag{93}$$

Applying this inequality to one half of the exponential factor in Equation (92) proves Equation (80). □

C.3 Rounds on which the optimal empirical mean is too small

Fix a deterministic number

$$\max_{j>1} \mu_j < b < \mu_1, \qquad \varepsilon_b := \mu_1 - b, \qquad d_b := b - \max_{j>1} \mu_j. \tag{94}$$

Both ε_b and d_b are positive.

Fix n for which $\sigma_n < \infty$ almost surely; this condition will be verified inductively in Proposition 4. Suppose that $\hat{\mu}_{1,n} \le b$ and write

$$\hat{\mu}_{1,n} = b - y, \qquad y \ge 0. \tag{95}$$

Using the threshold expected improvement from Equation (25) with $\varepsilon = \varepsilon_b$, define

$$e_n(y) := \mathrm{EI}_{\varepsilon_b}(n, b-y), \qquad E_n(y) := 2^{-(r+2)} e_n(y). \tag{96}$$

Equations (26)– (27) give

$$e_n(y) = \frac{1}{\sqrt{n}}\{\phi(\sqrt{n}\,y) - \sqrt{n}\,y\Phi(-\sqrt{n}\,y)\}, \tag{97}$$

$$e_n(y) \ge \frac{a_1}{\sqrt{n}(1+ny^2)} \exp\!\left(-\frac{ny^2}{2}\right). \tag{98}$$

Also $0 < E_n(y) < 1$.

Choose a constant $C_* > 0$, depending only on the fixed instance, K, M, and β, large enough for the estimates below, and define

$$u_n(y) := \max\left\{n+1, \left\lceil C_* \left(1 + B_{\mathrm{dev},-1} + \log\frac{8K}{E_n(y)}\right)\right\rceil\right\}. \tag{99}$$

We next verify all consequences required from this threshold. The elementary inequality

$$\log(em) \le 2\sqrt{m}, \qquad m \ge 1, \tag{100}$$

implies that, whenever $m \ge C_*(1 + B_{\mathrm{dev},-1})$,

$$\frac{B_{\mathrm{dev},-1} + \beta\log(em)}{m} \le \frac{1}{C_*} + \frac{2\beta}{\sqrt{C_*}}. \tag{101}$$

Choose C_* so that the right-hand side is at most $d_b^2/4$. Then every suboptimal arm with $N_j(t) \ge u_n(y)$ satisfies

$$\hat{\mu}_j(t) \le \mu_j + d_b/2 \le b - d_b/2. \tag{102}$$

Consequently, for its posterior sample,

$$\mathbb{P}(\theta_j > b \mid \widetilde{\mathcal{F}}_{t-1}) \le \exp\!\left(-\frac{N_j(t) d_b^2}{8}\right). \tag{103}$$

Let

$$c_* := \min\{d_b^2/8, \gamma_-^2/16\}.$$

By increasing C_* further so that

$$c_* C_* \ge C_\gamma + 1, \qquad A_\gamma e^{-c_* C_*} \le 1, \qquad c_* C_* \ge 1, \qquad e^{-d_b^2 C_*/8} \le \frac{1}{4r}, \tag{104}$$

Equations (80), (103), and (99) imply, for $N_i(t), N_j(t) \ge u_n(y)$ and distinct $i, j > 1$,

$$\min\{G_{ij}(t), G_{ji}(t)\} \le \frac{E_n(y)}{8K}, \tag{105}$$

$$\mathbb{P}(\theta_j > b \mid \widetilde{\mathcal{F}}_{t-1}) \le \frac{1}{4r}. \tag{106}$$

For example, the first inequality follows from

$$\begin{aligned} &A_\gamma e^{C_\gamma B_{\mathrm{dev},-1}} e^{-c_* u_n(y)} \\ &\quad \le A_\gamma e^{-c_* C_*} e^{-(c_* C_* - C_\gamma) B_{\mathrm{dev},-1}} \left(\frac{E_n(y)}{8K}\right)^{c_* C_*} \le \frac{E_n(y)}{8K}. \end{aligned}$$

The posterior-tail inequality follows in the same way, without the factor $A_\gamma e^{C_\gamma B_{\mathrm{dev},-1}}$; increasing C_* by a fixed amount makes its right-hand side at most $1/(4r)$.

For a round in the interval with $N_1(t) = n$, define the set

$$\mathcal{S}_n(t) := \{j > 1 : N_j(t) \geq u_n(y)\}. \tag{107}$$

If this set contains an arm with positive sampling probability, choose

$$\ell_t \in \arg\min\{\hat{\mu}_i(t) : i \in \mathcal{S}_n(t),\ \pi_{t,i} > 0\}. \tag{108}$$

Then, for every $j \in \mathcal{S}_n(t)$ with $\pi_{t,j} > 0$,

$$\hat{\mu}_{\ell_t}(t) \leq \hat{\mu}_j(t), \qquad G_{\ell_t j}(t) = \min\{G_{\ell_t j}(t), G_{j\ell_t}(t)\}. \tag{109}$$

The terms with $\pi_{t,j} = 0$ vanish. Equation (105) therefore gives

$$\sum_{j \in \mathcal{S}_n(t)} \pi_{t,j} G_{\ell_t j}(t) \leq \frac{E_n(y)}{8}. \tag{110}$$

We also need a bound that applies when either arm has not reached the threshold. For a posterior variable $\theta_i \sim \mathcal{N}(\hat{\mu}_i(t), 1/N_i(t))$,

$$\mathbb{E}[|\theta_i| \mid \widetilde{\mathcal{F}}_{t-1}] \leq |\hat{\mu}_i(t)| + 1. \tag{111}$$

For $j > 1$, Equations (70) and $\log(em)/m \leq 1$ give

$$|\hat{\mu}_j(t)| \leq |\mu_j| + \sqrt{B_{\mathrm{dev},-1}} + \sqrt{\beta}.$$

For arm 1 in the present interval, $|\hat{\mu}_1(t)| = |b - y| \leq |b| + y$. Hence, for every i, j,

$$G_{ij}(t) \leq \mathbb{E}[|\theta_i| + |\theta_j| \mid \widetilde{\mathcal{F}}_{t-1}] \tag{112}$$

$$\leq U_n(y, B_{\mathrm{dev},-1}) := C\{1 + y + \sqrt{1 + B_{\mathrm{dev},-1}}\}. \tag{113}$$

The purpose of the bound $G_{ij}(t) \leq U_n(y, B_{\mathrm{dev},-1})$ is to turn a lower bound on marginal EI into a lower bound on sampling probability. To see this, let $q_t := 1 - \pi_t(\mathcal{S}_n(t))$ be the probability of selecting arm 1 or an insufficiently sampled suboptimal arm. Whenever ℓ_t (defined in (108)) exists, the marginal EI is bounded by the weighted pairwise EI sum, so Equation (110) and the bound just obtained give

$$\bar{s}_{t,\ell_t} \leq \sum_{j=1}^{K} \pi_{t,j} G_{\ell_t j}(t) \leq \frac{E_n(y)}{8} + U_n(y, B_{\mathrm{dev},-1}) q_t.$$

If $q_t \leq 1/2$, the next lemma gives $\bar{s}_{t,1} \geq 2E_n(y)$, while EI balance gives $\bar{s}_{t,\ell_t} \geq \bar{s}_{t,1}$. The displayed upper bound then forces $q_t \geq E_n(y)/U_n(y, B_{\mathrm{dev},-1})$. If $q_t > 1/2$, a lower bound is already available. Thus the uniform upper bound on pairwise EI ensures a positive probability of either ending the current block by selecting arm 1, or bringing an insufficiently sampled arm closer to its threshold. We formalize this probability bound in Equation (120) and apply Lemma 14 with $u = u_n(y)$ and $p = p_n(y, B_{\mathrm{dev},-1})$ to obtain the expected block-length bound in Equation (129).

Lemma 13 (Lower bound on $\bar{s}_{t,1}$)**.** *If $\pi_t(\mathcal{S}_n(t)) \geq 1/2$, then*

$$\bar{s}_{t,1} \geq 2E_n(y). \tag{114}$$

Proof. Let $A_1, \ldots, A_r$ be the other $r = M-1$ virtual arm indices in the definition of $\bar{s}_{t,1}$. Conditional on $\widetilde{\mathcal{F}}_{t-1}$,

$$\mathbb{P}(A_1, \ldots, A_r \in \mathcal{S}_n(t)) = \pi_t(\mathcal{S}_n(t))^r \geq 2^{-r}. \tag{115}$$

Fix any realized indices $a_1, \ldots, a_r$ in this set. A union bound and Equation (106) yield

$$\mathbb{P}\left(\max_{1 \leq q \leq r} \theta_{a_q} \leq b \mid \widetilde{\mathcal{F}}_{t-1}, a_1, \ldots, a_r\right) \geq 1 - r\frac{1}{4r} = \frac{3}{4}. \tag{116}$$

Repeated indices use the same posterior coordinate, so the union bound remains valid. Arm 1 is independent of the posterior coordinates of the suboptimal arms. On the event in Equation (116),

$$\left(\theta_1 - \max_{1 \leq q \leq r} \theta_{a_q}\right)_+ \geq (\theta_1 - b)_+.$$

Using Equations (115), (116), and (96), we obtain

$$\begin{aligned}\bar{s}_{t,1} &\geq 2^{-r}\frac{3}{4}e_n(y)\\ &= 3\cdot 2^{-r-2}e_n(y) = 3E_n(y) \geq 2E_n(y).\end{aligned}$$

□

Combining the preceding estimates gives a lower bound on the probability of sampling arm 1 or a suboptimal arm whose count is below $u_n(y)$. If $\pi_t(\mathcal{S}_n(t)) < 1/2$, then this probability is larger than $1/2$. If $\pi_t(\mathcal{S}_n(t)) \geq 1/2$ and the set contains an arm with positive sampling probability, use ℓ_t from Equation (108). The EI balance condition, Equations (114), (110), and (113) give

$$2E_n(y) \leq \bar{s}_{t,1} \leq \bar{s}_{t,\ell_t} \tag{117}$$

$$\leq \sum_{j\in\mathcal{S}_n(t)} \pi_{t,j}G_{\ell_t j}(t) + U_n(y, B_{\mathrm{dev},-1})\left(\pi_{t,1} + \sum_{\substack{j>1\\ N_j(t)<u_n(y)}} \pi_{t,j}\right) \tag{118}$$

$$\leq \frac{E_n(y)}{8} + U_n(y, B_{\mathrm{dev},-1})\left(\pi_{t,1} + \sum_{\substack{j>1\\ N_j(t)<u_n(y)}} \pi_{t,j}\right). \tag{119}$$

If $\mathcal{S}_n(t)$ contains no arm with positive sampling probability, the probability in parentheses is one. Thus, in every case,

$$\pi_{t,1} + \sum_{\substack{j>1\\ N_j(t)<u_n(y)}} \pi_{t,j} \geq p_n(y, B_{\mathrm{dev},-1}) := \min\left\{\frac{1}{2}, \frac{E_n(y)}{U_n(y, B_{\mathrm{dev},-1})}\right\}. \tag{120}$$

We now prove the counting inequality used here and later.

Lemma 14 (Counting rounds until the next pull of arm 1)**.** *Fix an interval on which $N_1(t) = n$. Suppose there are an integrable, $\mathcal{G}_n$-measurable positive integer u and a $\mathcal{G}_n$-measurable number $p > 0$ such that, at every round in the interval,*

$$\pi_{t,1} + \sum_{\substack{j>1\\ N_j(t)<u}} \pi_{t,j} \geq p. \tag{121}$$

Then

$$\mathbb{E}[\sigma_{n+1} - \sigma_n \mid \mathcal{G}_n] \leq \frac{1 + (K-1)u}{p}. \tag{122}$$

Proof. For every deterministic $t \geq K+1$, define

$$V_t := \mathbf{1}\{\sigma_n \leq t < \sigma_{n+1}\}\left[1 + \sum_{j=2}^{K}(u - N_j(t))_+\right]. \tag{123}$$

At the start of the interval, $0 \leq V_{\sigma_n} \leq 1 + (K-1)u$. If $t < \sigma_{n+1}$ and arm 1 is sampled, then $t+1 = \sigma_{n+1}$ and $V_{t+1} = 0$, so $V_{t+1} - V_t \leq -1$. If a suboptimal arm with $N_j(t) < u$ is sampled, its summand decreases by one. Every other action leaves the bracketed quantity unchanged. Therefore

$$\mathbf{1}\{\sigma_n \leq t < \sigma_{n+1}\}(V_{t+1} - V_t) \leq -\mathbf{1}\{\sigma_n \leq t < \sigma_{n+1}\}\mathbf{1}\{I(t) = 1 \text{ or } [I(t) > 1, N_{I(t)}(t) < u]\}. \tag{124}$$

On $\{\sigma_n \leq t\}$, the $\mathcal{G}_n$-measurable variables u and p are also $\widetilde{\mathcal{F}}_{t-1}$-measurable. Using Equation (76) and then Equation (121),

$$\mathbb{E}\Big[\mathbf{1}\{\sigma_n \leq t < \sigma_{n+1}\}(V_{t+1} - V_t)\Big|\widetilde{\mathcal{F}}_{t-1}\Big] \leq -p\mathbf{1}\{\sigma_n \leq t < \sigma_{n+1}\}. \tag{125}$$

For a deterministic $Q \geq 1$, set $\rho_Q := \min\{\sigma_{n+1}, \sigma_n + Q\}$. The random-limit sum can be written with deterministic indices as

$$V_{\rho_Q} - V_{\sigma_n} = \sum_{t=K+1}^{\infty} \mathbf{1}\{\sigma_n \leq t < \rho_Q\}(V_{t+1} - V_t). \tag{126}$$

Apply the tower property to every term, first with the sum truncated at a deterministic upper limit and then by conditional dominated convergence. For fixed Q, the absolute value of the sum is at most $Q\{1 + (K-1)u\}$, which is finite and $\mathcal{G}_n$-measurable. The indicator $\mathbf{1}\{\sigma_n \leq t < \rho_Q\}$ is $\widetilde{\mathcal{F}}_{t-1}$-measurable. Since p is $\mathcal{G}_n$-measurable, Equation (125) gives

$$\mathbb{E}[V_{\rho_Q} - V_{\sigma_n} \mid \mathcal{G}_n] \leq -p\,\mathbb{E}\left[\sum_{t=K+1}^{\infty} \mathbf{1}\{\sigma_n \leq t < \rho_Q\}\middle| \mathcal{G}_n\right] \tag{127}$$

$$= -p\,\mathbb{E}[\rho_Q - \sigma_n \mid \mathcal{G}_n]. \tag{128}$$

Since $V_{\rho_Q} \geq 0$ and $V_{\sigma_n} \leq 1 + (K-1)u$,

$$\mathbb{E}[\rho_Q - \sigma_n \mid \mathcal{G}_n] \leq \frac{1 + (K-1)u}{p}.$$

Letting $Q \to \infty$ and applying monotone convergence proves Equation (122). □

On $\{\hat{\mu}_{1,n} \leq b\}$, set $y := b - \hat{\mu}_{1,n}$. Apply Lemma 14 with $u = u_n(y)$ and $p = p_n(y, B_{\mathrm{dev},-1})$. For fixed n, Equation (99) shows that $u_n(y)$ is bounded by a constant times $1 + B_{\mathrm{dev},-1} + ny^2 + \log(1 + ny^2)$; hence it is integrable by Equation (74) and the Gaussian law of $\hat{\mu}_{1,n}$. Since $1/p \leq 2 + U_n(y, B_{\mathrm{dev},-1})/E_n(y)$, we obtain, on $\{\hat{\mu}_{1,n} \leq b\}$,

$$\mathbb{E}[\sigma_{n+1} - \sigma_n \mid \mathcal{G}_n] \leq C\{1 + Ku_n(y)\}\left\{1 + \frac{U_n(y, B_{\mathrm{dev},-1})}{E_n(y)}\right\}. \tag{129}$$

For use in the logarithmic part of the proof, we also need a bound when $\hat{\mu}_{1,n} > b$. The next lemma supplies it without depending on n or on the value of $\hat{\mu}_{1,n}$.

Lemma 15 (Interval length when $\hat{\mu}_{1,n} > b$)**.** *There is a constant C_b such that*

$$\mathbf{1}\{\hat{\mu}_{1,n} > b\}\mathbb{E}[\sigma_{n+1} - \sigma_n \mid \mathcal{G}_n] \leq C_b(1 + B_{\mathrm{dev}})^2. \tag{130}$$

Proof. Let

$$a_b := \frac{1}{2}\left(b + \max_{j>1} \mu_j\right), \qquad c_b := b - a_b = \frac{d_b}{2} > 0.$$

If $\hat{\mu}_{1,n} > b$, Jensen's inequality gives

$$\mathbb{E}[(\theta_1 - a_b)_+ \mid \widetilde{\mathcal{F}}_{t-1}] \geq (\hat{\mu}_{1,n} - a_b)_+ \geq c_b. \tag{131}$$

Choose

$$u_b(B_{\mathrm{dev}}) := \lceil C_b(1 + B_{\mathrm{dev}}) \rceil, \tag{132}$$

where C_b is increased below when necessary. For a suboptimal arm with $N_j(t) \geq u_b(B_{\mathrm{dev}})$, the calculation in Equations (100)– (102), with a_b in place of b, gives

$$\hat{\mu}_j(t) \leq a_b - \frac{c_b}{2}, \qquad \mathbb{P}(\theta_j > a_b \mid \widetilde{\mathcal{F}}_{t-1}) \leq \frac{1}{4r}. \tag{133}$$

The exponential bound in Lemma 12, with the larger random variable B_{dev} in place of $B_{\mathrm{dev},-1}$, also gives

$$\min\{G_{ij}(t), G_{ji}(t)\} \leq \frac{2^{-r-3}c_b}{K} \tag{134}$$

for distinct suboptimal arms whose counts are both at least $u_b(B_{\mathrm{dev}})$. To see both claims explicitly, set $c_0 := c_b^2/8$. Their left-hand sides are bounded by

$$\exp\{-c_0 u_b(B_{\mathrm{dev}})\} \quad \text{and} \quad A_\gamma \exp\{C_\gamma B_{\mathrm{dev}} - \gamma_-^2 u_b(B_{\mathrm{dev}})/16\},$$

respectively. Taking C_b so that the coefficients of B_{dev} in both exponents are negative and their values at $B_{\mathrm{dev}} = 0$ are below the displayed constants proves Equations (133) and (134).

Define

$$\mathcal{S}_b(t) := \{j > 1 : N_j(t) \geq u_b(B_{\mathrm{dev}})\}.$$

If $\pi_t(\mathcal{S}_b(t)) \geq 1/2$, the calculation in Lemma 13, with a_b in place of b, gives

$$\bar{s}_{t,1} \geq 2^{-r}\frac{3}{4}c_b. \tag{135}$$

All posterior means are bounded by

$$\max_i |\mu_i| + \sqrt{B_{\mathrm{dev}}} + \sqrt{\beta},$$

so Equation (111) gives

$$\max_{i,j} G_{ij}(t) \leq C\sqrt{1 + B_{\mathrm{dev}}}. \tag{136}$$

If $\mathcal{S}_b(t)$ contains an arm with positive sampling probability, choose one, denoted by ℓ_t, whose empirical mean is the smallest among such arms. For every $j \in \mathcal{S}_b(t)$ with $\pi_{t,j} > 0$, $G_{\ell_t j}(t) = \min\{G_{\ell_t j}(t), G_{j\ell_t}(t)\}$; terms with $\pi_{t,j} = 0$ vanish. The EI balance condition and Equations (134), (135), and (136) therefore give

$$\begin{aligned} 3 \cdot 2^{-r-2}c_b \leq \bar{s}_{t,1} &\leq \bar{s}_{t,\ell_t} \\ &\leq \frac{2^{-r-3}c_b}{K} + C\sqrt{1 + B_{\mathrm{dev}}}\left(\pi_{t,1} + \sum_{\substack{j>1 \\ N_j(t)<u_b(B_{\mathrm{dev}})}} \pi_{t,j}\right). \end{aligned} \tag{137}$$

The difference between the first term and $2^{-r-3}c_b/K$ is a positive constant. Dividing Equation (137) by $C\sqrt{1 + B_{\mathrm{dev}}}$ yields

$$\pi_{t,1} + \sum_{\substack{j>1 \\ N_j(t)<u_b(B_{\mathrm{dev}})}} \pi_{t,j} \geq \frac{c}{\sqrt{1 + B_{\mathrm{dev}}}} \tag{138}$$

for a deterministic $c > 0$. If $\pi_t(\mathcal{S}_b(t)) < 1/2$, the left-hand side is larger than $1/2$. If the set contains no arm with positive sampling probability, the left-hand side equals one. Thus the same displayed lower bound holds in every case after decreasing c.

Lemma 14, applied with $u = u_b(B_{\mathrm{dev}})$ and $p = c/\sqrt{1 + B_{\mathrm{dev}}}$, now gives the following bound. The required integrability of $u_b(B_{\mathrm{dev}})$ follows from Equation (74).

$$\begin{aligned} \mathbf{1}\{\hat{\mu}_{1,n} > b\}\mathbb{E}[\sigma_{n+1} - \sigma_n \mid \mathcal{G}_n] &\leq C\{1 + Ku_b(B_{\mathrm{dev}})\}\sqrt{1 + B_{\mathrm{dev}}} \\ &\leq C_b(1 + B_{\mathrm{dev}})^{3/2} \\ &\leq C_b(1 + B_{\mathrm{dev}})^2, \end{aligned}$$

which proves the lemma. □

We can now sum the intervals on which arm 1 is underestimated.

Proposition 4 (Finite expected number of rounds with $\hat{\mu}_1(t) \leq b$)**.** *For every b satisfying Equation (94),*

$$\mathbb{E}_\mu\left[\sum_{t=K+1}^{\infty} \mathbf{1}\{\hat{\mu}_1(t) \leq b\}\right] < \infty. \tag{139}$$

Proof. First, every σ_n is finite almost surely. Indeed, $\sigma_1 = K + 1$. Conditional on $\mathcal{G}_n$, either $\hat{\mu}_{1,n} \leq b$, in which case Equation (129) is finite, or $\hat{\mu}_{1,n} > b$, in which case Equation (130) is finite. A nonnegative random variable with finite conditional expectation is finite almost surely. Induction on n therefore proves

$$\mathbb{P}(\sigma_n < \infty) = 1, \qquad n \geq 1. \tag{140}$$

On $\{\hat{\mu}_{1,n} \le b\}$, set $y := b - \hat{\mu}_{1,n}$ and use Equations (99), (98), and (113) in Equation (129). The lower bound on $E_n(y)$ first gives

$$\log \frac{8K}{E_n(y)} \le C + \frac{1}{2}\log n + \log(1+ny^2) + \frac{ny^2}{2}. \tag{141}$$

Using $\log n \le n$ and $\log(1+ny^2) \le ny^2$, we obtain the explicit bound

$$\begin{aligned} &\mathbf{1}\{\hat{\mu}_{1,n} \le b\}\mathbb{E}[\sigma_{n+1} - \sigma_n \mid \mathcal{G}_n] \\ &\quad \le C(1+B_{\mathrm{dev},-1})^{3/2} n^{5/2}(1+y)^5 \exp\!\left(\frac{ny^2}{2}\right). \end{aligned} \tag{142}$$

For completeness, the powers in this display follow from

$$\begin{aligned} u_n(y) &\le C\{n+1+B_{\mathrm{dev},-1} + \log n + \log(1+ny^2) + ny^2\} \\ &\le C(1+B_{\mathrm{dev},-1})n(1+y^2), \\ E_n(y)^{-1} &\le C\sqrt{n}(1+ny^2)e^{ny^2/2} \le Cn^{3/2}(1+y^2)e^{ny^2/2}, \\ U_n(y, B_{\mathrm{dev},-1}) &\le C(1+B_{\mathrm{dev},-1})^{1/2}(1+y), \end{aligned}$$

and $(1+y)(1+y^2)^2 \le (1+y)^5$.

Define

$$\mathcal{C}_n := \sigma(\hat{\mu}_{1,n}, X_{j,m} : j > 1,\ m \ge 1)\,. \tag{143}$$

This sigma-field is contained in $\mathcal{G}_n$ because all σ_n are finite and $\widetilde{\mathcal{F}}$ contains all reward sequences. Taking conditional expectation of Equation (142) down to $\mathcal{C}_n$ preserves the same right-hand side.

The variable $B_{\mathrm{dev},-1}$ is independent of $\hat{\mu}_{1,n} \sim \mathcal{N}(\mu_1, 1/n)$. Therefore

$$\begin{aligned} &\mathbb{E}[\mathbf{1}\{\hat{\mu}_{1,n} \le b\}(\sigma_{n+1} - \sigma_n)] \\ &\quad \le C\mathbb{E}[(1+B_{\mathrm{dev},-1})^{3/2}]n^{5/2}\int_0^\infty (1+y)^5 e^{ny^2/2}\sqrt{\frac{n}{2\pi}}e^{-n(\varepsilon_b+y)^2/2}\,\mathrm{d}y \qquad (144) \\ &\quad = C\mathbb{E}[(1+B_{\mathrm{dev},-1})^{3/2}]n^3 e^{-n\varepsilon_b^2/2}\int_0^\infty (1+y)^5 e^{-n\varepsilon_b y}\,\mathrm{d}y. \qquad (145) \end{aligned}$$

The equality uses the exact cancellation

$$-\frac{n}{2}(\varepsilon_b+y)^2 + \frac{ny^2}{2} = -\frac{n\varepsilon_b^2}{2} - n\varepsilon_b y. \tag{146}$$

Expanding $(1+y)^5$ and using

$$\int_0^\infty y^q e^{-n\varepsilon_b y}\,\mathrm{d}y = \frac{q!}{(n\varepsilon_b)^{q+1}}, \qquad q = 0,\dots,5, \tag{147}$$

gives

$$\int_0^\infty (1+y)^5 e^{-n\varepsilon_b y}\,\mathrm{d}y \le \frac{C_b}{n}. \tag{148}$$

Equations (74), (145), and (148) imply

$$\mathbb{E}[\mathbf{1}\{\hat{\mu}_{1,n} \le b\}(\sigma_{n+1} - \sigma_n)] \le C_b n^2 e^{-n\varepsilon_b^2/2}. \tag{149}$$

The right-hand side is summable in n. Finally, by Equation (77),

$$\sum_{t=K+1}^{\infty} \mathbf{1}\{\hat{\mu}_1(t) \le b\} = \sum_{n=1}^{\infty} \mathbf{1}\{\hat{\mu}_{1,n} \le b\}(\sigma_{n+1} - \sigma_n) \quad \text{almost surely.} \tag{150}$$

Taking expectations and summing Equation (149) proves the proposition. □

C.4 Asymptotic upper bound for a suboptimal arm

Fix a suboptimal arm $j > 1$, and choose

$$0 < \varepsilon < \min\left\{1, \frac{\Delta_{\min}}{5}\right\}, \qquad b_\varepsilon := \mu_1 - \varepsilon, \qquad d_{j,\varepsilon} := D(\mu_j + \varepsilon \| \mu_1 - \varepsilon) = \frac{(\Delta_j - 2\varepsilon)^2}{2}. \tag{151}$$

Fix $\eta > 0$ and choose $\kappa > 1$ so that

$$\rho_\kappa := \frac{\kappa}{\kappa + 1}, \qquad (1+\eta)\rho_\kappa > 1. \tag{152}$$

Define

$$m_T := \left\lceil \frac{r(1+\eta)\log T}{d_{j,\varepsilon}} \right\rceil. \tag{153}$$

On a round satisfying

$$\hat{\mu}_1(t) > b_\varepsilon, \qquad \hat{\mu}_j(t) \le \mu_j + \varepsilon, \tag{154}$$

the empirical gap is larger than $\Delta_j - 2\varepsilon$. The pairwise saturation inequality in Lemma 8 therefore gives

$$\pi_{t,j} \le A_{j,M} \exp\left\{ -\frac{d_{j,\varepsilon}}{r} \frac{N_1(t) N_j(t)}{N_1(t) + N_j(t)} \right\}. \tag{155}$$

Indeed, the exponent $f_t^\star / r$ in Equation (59) is at least the exponent displayed above. The constant $A_{j,M}$ is independent of t and T.

Lemma 10 and its proof show

$$\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{I(t) = j,\ \hat{\mu}_j(t) > \mu_j + \varepsilon\} \right] \le \frac{2}{\varepsilon^2}. \tag{156}$$

There are at most m_T pulls of arm j on rounds with $N_j(t) < m_T$. It remains to consider rounds satisfying Equations (154) and $N_j(t) \ge m_T$. First suppose

$$N_1(t) \ge \kappa N_j(t). \tag{157}$$

Then

$$\frac{N_1(t) N_j(t)}{N_1(t) + N_j(t)} \ge \rho_\kappa N_j(t) \ge \rho_\kappa m_T. \tag{158}$$

Equations (155) and (153) imply the following one-round bound. All conditions in the indicator are $\widetilde{\mathcal{F}}_{t-1}$-measurable, so

$$\begin{aligned} &\mathbb{E}[\mathbf{1}\{\hat{\mu}_1(t) > b_\varepsilon, \hat{\mu}_j(t) \le \mu_j + \varepsilon, N_j(t) \ge m_T, N_1(t) \ge \kappa N_j(t)\} \mathbf{1}\{I(t) = j\}] \\ &= \mathbb{E}[\mathbf{1}\{\hat{\mu}_1(t) > b_\varepsilon, \hat{\mu}_j(t) \le \mu_j + \varepsilon, N_j(t) \ge m_T, N_1(t) \ge \kappa N_j(t)\} \pi_{t,j}] \\ &\le A_{j,M} \exp\left(-\frac{d_{j,\varepsilon} \rho_\kappa m_T}{r} \right). \end{aligned} \tag{159}$$

Summing this inequality over at most T rounds gives

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{I(t) = j,\ \hat{\mu}_1(t) > b_\varepsilon,\ \hat{\mu}_j(t) \le \mu_j + \varepsilon,\ N_j(t) \ge m_T,\ N_1(t) \ge \kappa N_j(t)\} \right] \\ &\le A_{j,M} T \exp\left(-\frac{d_{j,\varepsilon} \rho_\kappa m_T}{r} \right) \qquad (160) \\ &\le A_{j,M} T^{1 - (1+\eta)\rho_\kappa} = o(1). \qquad (161) \end{aligned}$$

Now suppose $N_1(t) < \kappa N_j(t)$ and group the rounds by $n = N_1(t)$. On such a round, the harmonic count in Equation (155) satisfies

$$\frac{n N_j(t)}{n + N_j(t)} \ge \begin{cases} n/2, & 1 \le n \le m_T, \\ m_T/2, & m_T < n \le \kappa m_T, \\ n/(\kappa+1), & n > \kappa m_T. \end{cases} \tag{162}$$

The first line uses $N_j(t) \geq m_T \geq n$, the second uses $N_j(t) \geq m_T$ and $n > m_T$, and the third uses $N_j(t) > n/\kappa$.

Accordingly, define the deterministic bound

$$a_{n,T} := A_{j,M} \begin{cases} \exp\{-d_{j,\varepsilon} n/(2r)\}, & 1 \leq n \leq m_T, \\ \exp\{-d_{j,\varepsilon} m_T/(2r)\}, & m_T < n \leq \kappa m_T, \\ \exp\{-d_{j,\varepsilon} n/[r(\kappa+1)]\}, & n > \kappa m_T. \end{cases} \tag{163}$$

Equation (155) gives $\pi_{t,j} \leq a_{n,T}$ on every round in the nth interval under consideration.

For each n, let $A_{n,t}$ be the following pre-decision event:

$$\begin{aligned} A_{n,t} := \{K+1 \leq t \leq T,\ N_1(t) = n,\ \hat{\mu}_{1,n} > b_\varepsilon, \\ \hat{\mu}_j(t) \leq \mu_j + \varepsilon,\ N_j(t) \geq m_T,\ n < \kappa N_j(t)\}. \end{aligned} \tag{164}$$

Every condition in $A_{n,t}$ is $\widetilde{\mathcal{F}}_{t-1}$-measurable. Therefore Equations (76), (163), and the tower property give

$$\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{A_{n,t}\}\mathbf{1}\{I(t)=j\} \middle| \mathcal{G}_n\right] \tag{165}$$

$$= \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbf{1}\{A_{n,t}\}\pi_{t,j} \middle| \mathcal{G}_n\right] \tag{166}$$

$$\leq a_{n,T}\, \mathbf{1}\{\hat{\mu}_{1,n} > b_\varepsilon\}\mathbb{E}[\sigma_{n+1} - \sigma_n \mid \mathcal{G}_n] \tag{167}$$

$$\leq a_{n,T} C_{b_\varepsilon} (1 + B_{\mathrm{dev}})^2, \tag{168}$$

where the last line is Lemma 15. Taking expectations and using Equation (74), the total contribution from $N_1(t) < \kappa N_j(t)$ is at most a constant times $\sum_{n\geq 1} a_{n,T}$.

We now evaluate this sum. First,

$$\sum_{n=1}^{m_T} e^{-d_{j,\varepsilon} n/(2r)} \leq \frac{1}{e^{d_{j,\varepsilon}/(2r)} - 1} = O_{j,\varepsilon,M}(1). \tag{169}$$

Second, Equation (153) gives

$$\sum_{m_T < n \leq \kappa m_T} e^{-d_{j,\varepsilon} m_T/(2r)} \leq \kappa m_T e^{-d_{j,\varepsilon} m_T/(2r)} \tag{170}$$

$$\leq C \log T\, T^{-(1+\eta)/2} = o(1). \tag{171}$$

Third,

$$\sum_{n > \kappa m_T} e^{-d_{j,\varepsilon} n/[r(\kappa+1)]} \leq \frac{e^{-d_{j,\varepsilon}\kappa m_T/[r(\kappa+1)]}}{1 - e^{-d_{j,\varepsilon}/[r(\kappa+1)]}} \tag{172}$$

$$\leq C T^{-(1+\eta)\rho_\kappa} = o(1). \tag{173}$$

Thus the contribution in Equation (168), summed over n, is $O(1)$.

Combine Proposition 4, Equation (156), the at most m_T pulls before $N_j(t)$ reaches m_T, Equation (161), and Equations (169)– (173). Including the initial pull, we obtain

$$\mathbb{E}_\mu[N_j(T+1)] \leq m_T + O_{\mu,K,M,\varepsilon,\eta}(1) + o(1). \tag{174}$$

Divide by $\log T$ and use Equation (153):

$$\limsup_{T\to\infty} \frac{\mathbb{E}_\mu[N_j(T+1)]}{\log T} \leq \frac{r(1+\eta)}{d_{j,\varepsilon}}. \tag{175}$$

First let $\eta \downarrow 0$ and then let $\varepsilon \downarrow 0$. Since $d_{j,\varepsilon} \to D(\mu_j \| \mu_1)$, this proves

$$\limsup_{T\to\infty} \frac{\mathbb{E}_\mu[N_j(T+1)]}{\log T} \leq \frac{M-1}{D(\mu_j\|\mu_1)}. \tag{176}$$

Multiplying by Δ_j, summing over the finitely many suboptimal arms, and using $\mathrm{Reg}(T) = \sum_{j=2}^{K} \Delta_j N_j(T+1)$ proves Equation (9).

C.5 MATCHING LOWER BOUND FOR $M = 2$

Setting $M = 2$ in Equation (8) gives, for every $j > 1$,

$$\limsup_{T\to\infty} \frac{\mathbb{E}_\mu[N_j(T+1)]}{\log T} \le \frac{1}{D(\mu_j\|\mu_1)}.$$

The preceding logarithmic upper bound holds for every fixed pairwise-distinct unit-variance Gaussian instance. Hence ReMax with $M = 2$ is uniformly good: namely,

$$\mathbb{E}_\mu[\mathrm{Reg}(T)] = O_\mu(\log T) = o(T^a) \qquad \text{for every } a > 0$$

and for any distinct[4] set of arms. It is well-known that (Lai & Robbins, 1985), for any uniformly good algorithm, we have the following lower bound

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\mu[N_j(T+1)]}{\log T} \ge \frac{1}{D(\mu_j\|\mu_1)}, \qquad j > 1.$$

Combining the upper and lower bounds yields

$$\lim_{T\to\infty} \frac{\mathbb{E}_\mu[N_j(T+1)]}{\log T} = \frac{2}{\Delta_j^2}, \qquad j > 1.$$

Since

$$\mathbb{E}_\mu[\mathrm{Reg}(T)] = \sum_{j=2}^{K} \Delta_j \, \mathbb{E}_\mu[N_j(T+1)]$$

and the number of arms is finite, the preceding limit gives

$$\lim_{T\to\infty} \frac{\mathbb{E}_\mu[\mathrm{Reg}(T)]}{\log T} = 2\sum_{j=2}^{K} \frac{1}{\Delta_j}.$$

This completes the proof of Theorem 2.

D REMAXGRAD IMPLEMENTATION

ReMaxGrad (Nishimori et al., 2026) approximates the maximization in Equation (2) by stochastic gradient ascent in logit space. The outer loop corresponds to the bandit interaction rounds indexed by t. At each round t, with the posterior Π_{t-1} fixed, ReMaxGrad runs an inner optimization loop over the sampling distribution π. At inner optimization step ℓ, write $\pi = \mathrm{softmax}(\boldsymbol{z})$ and draw S independent posterior samples $\boldsymbol{\theta}_\ell^{(1)}, \ldots, \boldsymbol{\theta}_\ell^{(S)} \overset{\text{iid}}{\sim} \Pi_{t-1}$. The corresponding Monte Carlo objective is

$$\widehat{J}_{t,\ell}^{(M)}(\pi) := \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{A_{1:M}\sim\pi}\left[\max_{m\in[M]} \theta_{\ell,A_m}^{(s)}\right]. \tag{177}$$

For each posterior sample, the gradient of the inner expectation over the virtual actions is computed exactly; Monte Carlo sampling is used only for the posterior expectation. Let $\widehat{\boldsymbol{g}}_{t,\ell} := \nabla_\pi \widehat{J}_{t,\ell}^{(M)}(\pi)$. The gradient with respect to the logits and the sampled simplex KKT gap are

$$\widehat{\boldsymbol{h}}_{t,\ell} := \nabla_{\boldsymbol{z}} \widehat{J}_{t,\ell}^{(M)}(\pi) = \pi \odot (\widehat{\boldsymbol{g}}_{t,\ell} - \langle \pi, \widehat{\boldsymbol{g}}_{t,\ell}\rangle \mathbf{1}), \tag{178}$$

$$\widehat{\mathcal{G}}_{t,\ell}(\pi) := \max_{i\in[K]} \widehat{g}_{t,\ell,i} - \langle \pi, \widehat{\boldsymbol{g}}_{t,\ell}\rangle. \tag{179}$$

We use the sampled gap in Equation (179) as the early-stopping criterion. Algorithm 1 summarizes the complete procedure.

The first adaptive round starts from the uniform policy. In later rounds, the logits are warm-started from the preceding round, while the Adam (Kingma & Ba, 2014) moments are reset. Algorithm 1 is written for one trajectory; in the vectorized implementation, early stopping occurs only when the largest sampled gap in the batch is below ξ. Posterior draws and the sampled-gradient kernel use float32, whereas the posterior state, logits, Adam arithmetic, action probabilities, regret accumulation, and diagnostics use float64. Since the inner optimization is capped at $L = 50$, the implementation is a fixed-budget approximation to the exact ReMax optimizer.

[4]Note that excluding the case of non-distinct subclass of arms does not change the discussion of Lai & Robbins (1985).

Algorithm 1 ReMaxGrad

Require: Retry budget M, horizon T, and optimizer parameters S, L, η, ξ

Pull every arm once, update the posterior, and initialize $\boldsymbol{z} \leftarrow \mathbf{0}$
for $t = K+1, \ldots, T$ **do**
  Reset the Adam moments $\boldsymbol{m} \leftarrow \mathbf{0}$ and $\boldsymbol{v} \leftarrow \mathbf{0}$
  for $\ell = 1, \ldots, L$ **do**
    $\pi \leftarrow \mathrm{softmax}(\boldsymbol{z})$
    Draw S fresh posterior samples from Π_{t-1}
    Compute $\widehat{\boldsymbol{g}}_{t,\ell}$ and $\widehat{\mathcal{G}}_{t,\ell}(\pi)$
    if $\widehat{\mathcal{G}}_{t,\ell}(\pi) \leq \xi$ **then**
      break
    end if
    Compute $\widehat{\boldsymbol{h}}_{t,\ell}$ from Equation (178)
    Update $(\boldsymbol{z}, \boldsymbol{m}, \boldsymbol{v})$ by Adam ascent with learning rate η
  end for
  Set $\pi_t \leftarrow \mathrm{softmax}(\boldsymbol{z})$ and sample $I(t) \sim \pi_t$
  Observe the reward and update the posterior to obtain Π_t
end for

Table 1: ReMaxGrad settings used in the main experiments.

Parameter	Setting
Retry budget M	$\{2, 3, 4\}$
Posterior samples per gradient step S	50
Maximum Adam steps per round L	50
Learning rate η	0.02
Adam $(\beta_1, \beta_2, \epsilon_{\mathrm{Adam}})$	$(0.9, 0.999, 10^{-8})$
Sampled KKT tolerance ξ	10^{-7}
Batch size (fixed-instance experiment)	100
Batch size (random-instance experiment)	1,000

Comparison with exact ReMax for $M = 2$ For $M = 2$, define $Q_{t,ij} := \mathbb{E}[\max\{\theta_i, \theta_j\}]$ for $i \neq j$ and $Q_{t,ii} := \mathbb{E}[\theta_i] = \hat{\mu}_i(t)$. The ReMax objective is then the concave quadratic program

$$J_t^{(2)}(\pi) = \pi^\top Q_t \pi, \qquad \pi \in \Delta^{K-1}. \tag{180}$$

Let $\mathcal{A} := \mathrm{supp}(\pi_t^\star)$. The KKT conditions give $(Q_t\pi_t^\star)_i = \lambda$ for $i \in \mathcal{A}$ and $(Q_t\pi_t^\star)_i \leq \lambda$ for $i \notin \mathcal{A}$. Hence, whenever $Q_{t,\mathcal{A}\mathcal{A}}$ is invertible,

$$\pi_{t,\mathcal{A}}^\star = \frac{Q_{t,\mathcal{A}\mathcal{A}}^{-1}\mathbf{1}}{\mathbf{1}^\top Q_{t,\mathcal{A}\mathcal{A}}^{-1}\mathbf{1}}, \qquad \pi_{t,\mathcal{A}^c}^\star = \mathbf{0}. \tag{181}$$

The analytical solver updates the active set until positivity and the inactive-arm inequalities are satisfied. To measure the optimization error of ReMaxGrad against this exact objective, we use the KKT gap

$$\mathcal{G}_t^{(2)}(\pi) := \max_{i\in[K]} [2Q_t\pi]_i - \langle \pi, 2Q_t\pi \rangle, \tag{182}$$

which is zero at the exact maximizer and upper-bounds $J_t^{(2)}(\pi_t^\star) - J_t^{(2)}(\pi)$ by concavity (Vandenberghe & Boyd, 2004).

We compare the analytical $M = 2$ optimizer with ReMaxGrad on the fixed instance in Section 4.2, using ten batches of 100 trajectories with $T = 20{,}000$. ReMaxGrad uses the same optimizer settings and batch size as the main fixed-instance experiment (Table 1). We evaluate its KKT gap under the exact posterior objective in Equation (182), rather than the sampled objective.

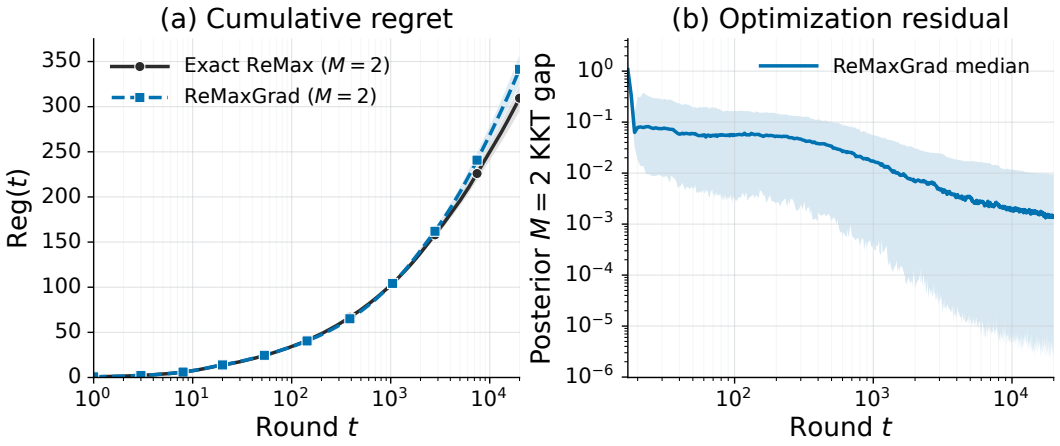

Figure 6: Comparison of the analytical $M = 2$ optimizer and ReMaxGrad on the fixed instance ($T = 20{,}000$, $\xi = 10^{-7}$). (a) Mean cumulative regret over 1,000 trajectories, with one-standard-error bands. (b) Median exact-posterior KKT gap of ReMaxGrad, with 5th–95th percentile bands.

Results Figure 6 shows that ReMaxGrad follows the overall regret trend of the analytical optimizer, with a higher mean cumulative regret at $T = 20{,}000$ (341.4 versus 309.4). Its median exact-posterior KKT gap generally decreases over time, reaching approximately 1.4×10^{-3} at the final round. In contrast, the exact-posterior KKT gap of the analytical solver remains below 10^{-8} on every adaptive round.