Posterior Tracking for Best-Arm Identification in Bernoulli Bandits

Kaito Ariu
CyberAgent
Tohoku University
kaito_ariu@cyberagent.co.jp

Junpei Komiyama
Mohamed bin Zayed University of Artificial Intelligence
RIKEN AIP
junpei@komiyama.info

Abstract

We study fixed-confidence best-arm identification in Bernoulli bandits. A learner sequentially samples from K arms and aims to identify the unique best arm while keeping the error probability at most δ. We ask whether the classical lower bound on expected sample complexity can be matched asymptotically without ongoing forced exploration. We propose a posterior-tracking algorithm that first samples each arm once, then repeatedly draws a bandit instance from the posterior, computes the oracle allocation for that instance, and randomly selects the next arm according to that allocation. Stopping is based on a Bernoulli generalized likelihood ratio test. Since the algorithm has no explicit forced-exploration mechanism after initialization, the main technical challenge is to show that posterior randomness alone sustains sufficient exploration. We prove that, with summable tail probabilities, every arm is sampled at least logarithmically often. This yields an integrable stabilization time after which the true best arm remains the empirical leader and the corresponding generalized likelihood ratio statistic grows linearly at the information-theoretic lower-bound rate. Consequently, the proposed procedure identifies the best arm with error probability at most δ and achieves asymptotically optimal expected sample complexity.

1 Introduction

Best-arm identification (BAI) is a central problem in pure exploration for multi-armed bandits. In the fixed-confidence setting, a learner sequentially samples from K arms and aims to identify the unique best arm while keeping the error probability at most a given value δ. The sequential-design viewpoint, which combines adaptive sampling and stopping, traces back to the framework of Chernoff (1959). For one-parameter exponential families, and in particular for Bernoulli bandits, Garivier and Kaufmann (2016) characterized the instance-dependent information-theoretic lower bound on the expected stopping time and introduced Track-and-Stop, which matches this lower bound asymptotically.

The central question of this paper is whether asymptotic optimality can be achieved without an explicit rule that forces additional pulls of under-sampled arms after initialization. In place of such a rule, we consider exploration driven by posterior randomness. Specifically, we ask:

Can posterior-driven sampling achieve asymptotic optimality without forced exploration?

1.1 Contributions

We answer this question affirmatively for posterior tracking in fixed-confidence best-arm identification over K-armed Bernoulli bandits. The sampling rule proceeds as follows: rather than tracking a deterministic target allocation derived from the unknown true parameter, each round draws a bandit

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

instance from the posterior (i.e., a posterior sample of the arm-mean vector), computes its oracle allocation, and samples according to that allocation. Stopping is performed by a Bernoulli generalized likelihood ratio (GLR) rule, following the classical sequential design and sequential-GLR lineage of Chernoff (1959), Garivier and Kaufmann (2016), and Kaufmann and Koolen (2021).

The main difficulty is to prove that this posterior randomization alone provides sufficient exploration. Since no forced exploration is imposed, it is not a priori clear that suboptimal arms will continue to be sampled sufficiently often, nor that the allocations induced by posterior draws will stabilize fast enough to support optimal stopping. Our analysis resolves this difficulty in several steps. First, we prove a logarithmic exploration property with summable tails, showing that every arm is sampled at least logarithmically often. Second, we establish an integrable stabilization time after which the true best arm remains the empirical leader and its GLR statistic grows linearly at the information-theoretic lower-bound rate. Third, combining this stabilization result with an anytime threshold for the GLR statistic yields both δ-correctness and a lower-bound-matching asymptotic upper bound on the expected sample complexity. Together with the classical lower bound of Garivier and Kaufmann (2016), this implies asymptotic optimality of the proposed procedure.

Our proof differs from classical tracking analyses in the first step, where we establish logarithmic exploration without a forced-exploration rule. The exploration argument is inspired by the long-gap argument of Kaufmann et al. (2012) for Thompson sampling: for each arm, a long gap contains a long block on which the posterior parameters of under-sampled arms remain fixed, and anti-concentration for these arms, posterior concentration for the saturated arms, and the oracle-allocation lower bound make it unlikely that the algorithm avoids the under-sampled arms throughout the block.

Our results show that posterior tracking provides an alternative to tracking rules that rely on explicit forced exploration in fixed-confidence Bernoulli BAI. Conceptually, they show that posterior sampling can be combined with oracle-allocation tracking within a single asymptotically optimal procedure.

1.2 Related work

We focus here on the closest posterior-randomized and top-two approaches; additional related work on classical confidence-bound methods, lower-bound tracking, game formulations, Thompson-style pure exploration, and related pure-exploration settings is deferred to Appendix A.

Top-two methods are a prominent line of posterior-randomized algorithms for best-arm identification. Russo (2020) introduced simple Bayesian top-two rules, and subsequent work studied top-two expected improvement, top-two Thompson sampling (TTTS), top-two transportation cost (T3C), and general top-two schemes (Qin et al., 2017; Shang et al., 2020; Jourdan et al., 2022). These methods randomize between a leader and a challenger, often using a fixed mixing parameter β. Existing fixed-confidence analyses of TTTS and T3C establish β-optimality: they match the best constant subject to allocating an asymptotic fraction β to the true best arm, but a fixed β need not match the instance-optimal lower-bound constant.

Recent refinements adapt the leader–challenger choice more carefully. You et al. (2023) introduce information-directed selection, derived from dual optimal-allocation conditions, which dynamically chooses between the top-two candidates according to their information gain. Combined with top-two Thompson sampling, this yields asymptotic optimality for Gaussian best-arm identification without explicit forced exploration; the broader dual-directed framework of Qin and You (2025) extends this principle to richer pure-exploration goals such as thresholding bandits and ϵ-best-arm identification. Lee et al. (2024) combine Thompson sampling with a best-challenger rule for fixed-confidence BAI in canonical one-parameter exponential families, showing that Thompson-style randomness can explore without explicit forcing and without solving a full allocation problem at every round. Their exact asymptotic optimality guarantee is for two-armed problems, while for general K they establish near-optimality. The optimal top-two algorithms of Bandyopadhyay et al. (2024) attain asymptotic optimality for general K, but their theoretical sampling rules include an explicit forced-exploration step that samples under-explored arms with fewer than N^{α} pulls, where N denotes the current total number of pulls in their notation. Thus their asymptotic optimality guarantees rely on an explicit forced-exploration rule with a growing sample-count threshold.

These works show that posterior and Thompson-style randomization can be effective for best-arm identification. Their exact optimality guarantees, however, either apply to Gaussian or two-armed regimes, or rely on explicit forced exploration in the general-K case; moreover, the sampling

rules remain leader–challenger top-two designs. In contrast, we study ordinary K-armed Bernoulli BAI in the frequentist fixed-confidence setting and prove exact characteristic-time optimality in expected stopping time by tracking oracle allocations induced by posterior draws, without an explicit forced-exploration rule after initialization.

Murphy Sampling (Kaufmann et al., 2018) is another closely related posterior-randomized design. It studies sequential testing of whether the minimum arm mean lies below a threshold and gives an anytime posterior-sampling rule without δ-dependent forced exploration. Its setting and guarantee differ from ours: the problem is lowest-mean threshold testing rather than ordinary best-arm identification, and its asymptotic stopping-time guarantee is stated in almost-sure terms rather than as an expected sample-complexity guarantee for ordinary Bernoulli BAI.

The closest posterior-tracking work is Tabata et al. (2023), who study classification bandits. Their threshold-based classification structure leads to a different oracle geometry from ordinary best-arm identification. In contrast, we prove that tracking oracle allocations induced by posterior draws is δ-correct and achieves the characteristic-time constant of Garivier and Kaufmann (2016) in expected stopping time for ordinary K-armed Bernoulli BAI.

The remainder of the paper is organized as follows. Section 2 introduces the Bernoulli BAI setting, the posterior model, and the oracle quantities. Section 3 presents the posterior-tracking algorithm and the GLR stopping rule. Section 4 states the main guarantees and the key technical ingredients. Section 5 explains why posterior tracking explores enough and isolates the deterministic and probabilistic ingredients behind stabilization. Section 6 summarizes how stabilization yields the fixed-confidence guarantee and the asymptotically optimal expected sample complexity bound; full proofs are given in Appendix D.

2 Setting

Let $K \geq 2$ be the number of arms. Pulling arm $i \in \{1, \dots, K\}$ produces a Bernoulli reward $X \sim \mathrm{Bern}(\mu_i)$ with an unknown mean $\mu_i \in (0,1)$. Write $\mu = (\mu_1, \dots, \mu_K) \in (0,1)^K$ and assume that the best arm

$$a^\star(\mu) = \operatorname*{arg\,max}_{1 \leq i \leq K} \mu_i$$

is unique. For $t \geq 1$, let I_t be the chosen arm and let $H_t := (I_1, X_1, \dots, I_t, X_t)$ denote the history. Let $(\mathcal{F}_t)_{t \geq 0}$ be the natural filtration:

$$\mathcal{F}_t := \sigma(H_t) = \sigma(I_1, X_1, \dots, I_t, X_t), \qquad t \geq 1,$$

with $\mathcal{F}_0$ the trivial sigma-field. Define the pull counts and empirical means by

$$N_i(t) = \sum_{s=1}^{t} \mathbf{1}\{I_s = i\}, \qquad S_i(t) = \sum_{s=1}^{t} \mathbf{1}\{I_s = i\} X_s, \qquad \hat{\mu}_i(t) = \frac{S_i(t)}{N_i(t)} \ \ (N_i(t) \geq 1).$$

2.1 Beta posterior and oracle quantities

Assume that the arm means have independent Beta priors

$$\mu_i \sim \mathrm{Beta}(\alpha_i, \beta_i), \qquad \alpha_i, \beta_i \in \mathbb{N},\ \alpha_i \geq 1,\ \beta_i \geq 1,$$

across i. Then the posterior factorizes and, for each arm i,

$$\mu_i \mid \mathcal{F}_t \sim \mathrm{Beta}(\alpha_i + S_i(t),\ \beta_i + N_i(t) - S_i(t)). \tag{1}$$

We denote the posterior at time t by Π_t and a posterior draw by $\theta_{t+1} \sim \Pi_t$. Let Δ_K denote the probability simplex:

$$\Delta_K = \{w \in \mathbb{R}^K :\ w_i \geq 0,\ \sum_{i=1}^{K} w_i = 1\}.$$

Throughout, for any $x \in (0,1)^K$, define

$$a^\star(x) := \min \operatorname*{arg\,max}_{1 \leq i \leq K} x_i.$$

Thus ties, when present, are broken by the smallest index. For $p, q \in (0,1)$ define the Bernoulli KL divergence $d(p,q) = \mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$. When either argument of kl is equal to 0 or 1, we use the standard boundary conventions collected in Appendix C. For $x \in (0,1)^K$ and $w \in \Delta_K$, define

$$F(x,w) := \inf_{\lambda \in \mathrm{Alt}(x)} \sum_{i=1}^{K} w_i \, d(x_i, \lambda_i) \text{ with } \mathrm{Alt}(x) := \Big\{ \lambda \in (0,1)^K : \exists j \neq a^\star(x) \text{ s.t. } \lambda_j \geq \lambda_{a^\star(x)} \Big\}.$$

Let $w^\star(x) \in \arg\max_{w \in \Delta_K} F(x,w)$ be a maximizer. For the true instance μ, define the complexity term

$$T^\star(\mu) := \Big(\max_{w \in \Delta_K} F(\mu, w) \Big)^{-1}.$$

Appendix C collects basic KL conventions, binomial and beta tail bounds, posterior construction details, and continuity properties of the oracle allocation.

3 Algorithm

Algorithm 1 summarizes the posterior-tracking rule together with the GLR stopping condition. Here Π_{t-1} denotes the posterior law of μ conditional on $\mathcal{F}_{t-1}$; by (1), it is the product of the armwise Beta posteriors, and $\theta_t \sim \Pi_{t-1}$ is a posterior draw at round t.

Algorithm 1 Posterior-tracking with GLR stopping

Require: Confidence level $\delta \in (0,1)$
Pull each arm once
for $t = K+1, K+2, \ldots$ **do**
  Draw $\theta_t \sim \Pi_{t-1}$
  Compute $w_t := w^\star(\theta_t) \in \Delta_K$
  Sample $I_t \mid (\mathcal{F}_{t-1}, \theta_t) \sim \mathrm{Cat}(w_t)$ and observe X_t
  Update $(N_i(t), \hat{\mu}_i(t))_{i=1}^K$
  Compute $(\Gamma_i(t))_{i=1}^K$ and let J_t be the smallest index attaining $\max_{1 \leq i \leq K} \Gamma_i(t)$
  if $\Gamma_{J_t}(t) \geq \beta(t, \delta)$ **then**
    return $\hat{a} := J_t$ and stop at $\tau_\delta := t$
  end if
end for

3.1 Posterior-tracking sampling rule

After the one-sample-per-arm initialization (so $N_i(K) = 1$ for all i), each round $t > K$ draws $\theta_t \sim \Pi_{t-1}$, computes $w_t := w^\star(\theta_t) \in \Delta_K$, and samples $I_t \mid (\mathcal{F}_{t-1}, \theta_t) \sim \mathrm{Cat}(w_t)$, where $\mathrm{Cat}(w_t)$ denotes the categorical distribution on $\{1, \ldots, K\}$ with probability vector w_t; see Algorithm 1 for the full loop.

3.2 GLR statistic and stopping rule

For each arm i, define the generalized likelihood ratio statistic

$$\Gamma_i(t) = \inf_{\lambda \in \mathrm{Alt}_i} \sum_{k=1}^{K} N_k(t) \, \mathrm{kl}(\hat{\mu}_k(t), \lambda_k), \qquad \mathrm{Alt}_i := \Big\{ \lambda \in (0,1)^K : \exists j \neq i \text{ s.t. } \lambda_j \geq \lambda_i \Big\},$$

and let J_t be the smallest index attaining $\max_{1 \leq i \leq K} \Gamma_i(t)$, so ties are broken deterministically by the smallest index.

Lemma 1 (Closed-form Bernoulli GLR statistics)**.** *Assume $N_i(t) \geq 1$ for all i. For $i \neq j$ define*

$$\hat{m}_{i,j}(t) := \frac{N_i(t)\hat{\mu}_i(t) + N_j(t)\hat{\mu}_j(t)}{N_i(t) + N_j(t)},$$
$$Z_{i,j}(t) := N_i(t)\mathrm{kl}(\hat{\mu}_i(t), \hat{m}_{i,j}(t)) + N_j(t)\mathrm{kl}(\hat{\mu}_j(t), \hat{m}_{i,j}(t)).$$

Then

$$\Gamma_i(t) = \begin{cases} 0, & \hat{\mu}_i(t) < \max_{k \neq i} \hat{\mu}_k(t), \\ \min_{j \neq i} Z_{i,j}(t), & \hat{\mu}_i(t) = \max_k \hat{\mu}_k(t). \end{cases}$$

The full proof is given in Appendix C.5.1.

For a confidence parameter $\delta \in (0,1)$ and a threshold $\beta(t,\delta)$ specified later in Proposition 2, define the stopping time and recommendation by

$$\tau_\delta := \inf\Big\{ t \geq K+1 : \ \Gamma_{J_t}(t) \geq \beta(t,\delta) \Big\}, \qquad \hat{a} := J_{\tau_\delta}.$$

Because $\beta(t,\delta) > 0$, Lemma 1 implies that stopping can occur only when J_t is the unique empirical best arm.

4 Main Results

This section states the main guarantees for the proposed sampling rule. The first two results give fixed-confidence correctness and the asymptotic expected-sample-complexity upper bound. The remaining two results provide the technical basis for the analysis: every arm is sampled at least logarithmically often with summable tail probabilities, and the corresponding stabilization time is integrable.

4.1 Correctness and sample-complexity guarantees

Theorem 1 (δ-correctness)**.** *For every $\delta \in (0,1)$, the recommendation $\hat{a}$ returned by the posterior-tracking algorithm at time τ_δ satisfies*

$$\mathbb{P}_\mu(\hat{a} \neq a^\star(\mu)) \leq \delta.$$

Theorem 2 (Asymptotic expected-sample-complexity upper bound)**.** *Fix the threshold $\beta(t,\delta)$ from Proposition 2 and the stopping time τ_δ above. Then*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\log(1/\delta)} \leq T^\star(\mu).$$

Combined with the classical lower bound of Garivier and Kaufmann (2016), Theorem 2 yields asymptotic optimality.

4.2 From logarithmic exploration to stabilization

Proposition 1 (Summable logarithmic exploration)**.** *There exists a constant $A_0 = A_0(\mu) \geq 1$ such that for any $A \geq A_0$ there exist constants $C = C(\mu, A) > 0$, $c = c(\mu, A) > 1$, and $t_0 = t_0(\mu, A) \geq 3$ such that for all $t \geq t_0$,*

$$\mathbb{P}_\mu\Big(\min_{1 \leq i \leq K} N_i(t) < A \log t \Big) \leq C\, t^{-c}. \qquad \text{(S-LE)}$$

Theorem 3 (Integrable stabilization)**.** *Assume that $\mu \in (0,1)^K$ has a unique best arm. Fix $\alpha \in (0,1)$ and define*

$$t_\alpha := \inf\Big\{ t \geq K+1 : \ J_s = a^\star(\mu) \ \text{ and } \ \Gamma_{a^\star(\mu)}(s) \geq \frac{1-\alpha}{T^\star(\mu)}\, s \ \text{ for all } s \geq t \Big\}.$$

Then $\mathbb{E}_\mu[t_\alpha] < \infty$.

Proposition 1 ensures that each arm is sampled at least logarithmically often, with summable tail probabilities. This logarithmic sampling lower bound is used to prove posterior concentration and control of the empirical allocation, which lead to the integrable stabilization statement in Theorem 3. The full proofs of Proposition 1 and Theorem 3 are deferred to Appendices F.1 and E, respectively. Section 5 gives the main proof ideas for these two technical statements, while Section 6 explains how they imply Theorems 1 and 2; full proofs of the final guarantees are given in Appendix D.

5 Sufficient Exploration under Posterior Tracking

This section gives the proof structure behind the two technical statements in Section 4. The main point is that posterior tracking samples every arm logarithmically often even without any explicit rule that forces pulls of under-sampled arms. We first sketch the proof of Proposition 1, then show how accurate empirical means and allocation proportions imply stabilization, and finally combine these ingredients to obtain the integrable stabilization statement in Theorem 3. The full proofs of the two technical statements are deferred to the appendix.

5.1 Summable logarithmic exploration

Proof sketch of Proposition 1. We build on the long-gap argument used by Kaufmann et al. (2012) in their finite-time analysis of Thompson sampling with Bernoulli rewards, and adapt it to posterior tracking. Fix an arm i and a horizon t. If $N_i(t) < A \log t$, then the counting argument in Appendix F.1 shows that one of the inter-pull gaps of arm i, including the terminal gap, has length

$$L \asymp \frac{t}{A \log t}.$$

For such a gap interval I, an interruption is a round in I at which the algorithm pulls an arm $k \neq i$ that is not saturated, i.e., $N_k(s-1) < A \log t$. Since each interruption increases the count of some arm $k \neq i$ until that arm becomes saturated, the number of interruptions is at most

$$M = (K-1)\lceil A \log t \rceil.$$

Hence the deterministic reduction in Lemma 9 extracts a non-interruption block $\widetilde{B}_q$ of length

$$L' \gtrsim \frac{t}{KA^2(\log t)^2}.$$

Let T_q be the time just before this block, so that $\widetilde{B}_q = \{T_q + 1, \dots, T_q + L'\}$, let $\mathcal{S}_{T_q}$ be the saturated set, and set $\mathcal{U} = \{1, \dots, K\} \setminus \mathcal{S}_{T_q}$. On the event Ξ_ℓ, the first ℓ rounds after T_q sample only arms in $\mathcal{S}_{T_q}$; consequently the posterior parameters of the arms in $\mathcal{U}$ remain equal to their values at time T_q. Lemma 11 then gives conditional independence of the current posterior draws of the arms in $\mathcal{U}$, together with an anti-concentration bound at scale $\delta_t = (\log t)^{-4}$.

The probability of $\Xi_{L'}$ is controlled by separating the rounds in $\widetilde{B}_q$ according to the saturated-good event

$$\mathrm{Good}_s := \bigcap_{k \in \mathcal{S}_{T_q}} \{|\theta_{k,s} - \mu_k| \leq r_\mu\}.$$

Lemma 8 implies that the number $\mathrm{Bad} = \sum_{s \in \widetilde{B}_q} \mathbf{1}\{\mathrm{Good}_s^c\}$ exceeds $L'/2$ with probability at most of order t^{-Ac_0}. On the complementary event $\{\mathrm{Bad} \leq L'/2\}$, at least half of the block satisfies Good_s. For such rounds, the escape probability from $\mathcal{S}_{T_q}$ is bounded from below. If the true best arm belongs to $\mathcal{S}_{T_q}$, the anti-concentration estimate for the arms in $\mathcal{U}$ and the deterministic oracle-allocation lower bound in Lemma 12 imply

$$\sum_{u \in \mathcal{U}} w_u^\star(\theta_s) \geq \frac{1}{2K(\log t)^9}$$

with conditional probability bounded away from zero. Iterating the resulting one-step bound over the rounds satisfying Good_s gives

$$\mathbb{P}_\mu\big(\Xi_{L'} \cap \{\mathrm{Bad} \leq L'/2\} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\}\big) \leq \exp\left\{-c\,\frac{L'}{(\log t)^9}\right\}.$$

It remains to handle the case $a^\star(\mu) \in \mathcal{U}$. We treat this case in three steps. First, conditional on $\mathcal{F}_{T_q}$, the same one-step escape argument uses the conditional posterior probability

$$\Phi_q := \mathbb{P}_\mu\Big(\theta_{a^\star(\mu),T_q+1} \geq \mu^-_{a^\star(\mu)} \,\Big|\, \mathcal{F}_{T_q}\Big), \qquad \mu^-_{a^\star(\mu)} := \mu_{a^\star(\mu)} - \Delta_{\min}/4.$$

This yields a conditional escape probability of at least $\Phi_q/[8K(\log t)^9]$ at every good round as long as no arm in $\mathcal{U}$ has been selected. Second, on the event $\mathcal{M}_q$, Lemma 14 lower-bounds Φ_q by a

polynomial factor, using the Beta lower-tail estimate in Lemma 13; iteration over the block then gives a super-polynomially small bound on the probability of remaining in $\mathcal{S}_{T_q}$ throughout the block. Third, still in the case $a^\star(\mu) \in \mathcal{U}$, the complementary event $\mathcal{M}_q^c$ is controlled by a Chernoff bound together with the fact that $R_\mu > 1$, yielding a summable polynomial bound. Combining these two bounds with the saturated-tail estimate for Bad and the preceding bound for the case $a^\star(\mu) \in \mathcal{S}_{T_q}$ gives

$$\mathbb{P}_\mu(\widetilde{E}_{i,j,q}) \le Ct^{-\gamma} \qquad \text{for some } \gamma > 1,$$

which is Lemma 10. Finally, a union bound over arms i, gap indices j, and block indices q introduces only logarithmic factors, which are absorbed into the exponent. This proves the summable logarithmic exploration bound in Proposition 1. A full proof is deferred to Appendix F.1.

5.2 From accurate allocations to stabilization

Fix $\alpha \in (0,1)$ and write $a^\star := a^\star(\mu)$ and $w^\star := w^\star(\mu)$. By Lemma 4 in Appendix E, applied with $\varepsilon = \alpha$, there exist $\rho > 0$ and $\eta > 0$ such that, whenever

$$\|\hat{\mu}(t) - \mu\|_\infty \le \rho \quad \text{and} \quad \|N(t)/t - w^\star\|_\infty \le \eta,$$

one has

$$J_t = a^\star \quad \text{and} \quad \Gamma_{a^\star}(t) \ge \frac{1-\alpha}{T^\star(\mu)}\, t.$$

This deterministic implication reduces stabilization to accurate empirical means and accurate allocation proportions.

5.3 Integrable stabilization time

Theorem 3 follows by combining the deterministic implication above with two random stabilization times. Define the allocation stabilization time

$$\tau_A := \inf\Big\{ t \ge K : \|N(s)/s - w^\star\|_\infty \le \eta \text{ for all } s \ge t \Big\},$$

and the time immediately after the last empirical failure while the allocation condition holds,

$$\tau_B := 1 + \sup\Big\{ s \ge K : \|\hat{\mu}(s) - \mu\|_\infty > \rho,\ \|N(s)/s - w^\star\|_\infty \le \eta \Big\},$$

with the convention $\sup \emptyset := K$. By the definition of τ_B, for every integer $s \ge \tau_B$, the allocation condition $\|N(s)/s - w^\star\|_\infty \le \eta$ implies $\|\hat{\mu}(s) - \mu\|_\infty \le \rho$. Together with the definition of τ_A and the deterministic implication above, this gives

$$t_\alpha \le 1 + \max\{\tau_A, \tau_B\}.$$

To control τ_B, note that on the event $\{\|N(s)/s - w^\star\|_\infty \le \eta\}$ every arm has already been sampled linearly often, so Hoeffding bounds make empirical-mean failures exponentially rare in s; summing the resulting tails gives $\mathbb{E}_\mu[\tau_B] < \infty$. To control τ_A, write $w_t := w^\star(\theta_t)$ and decompose

$$N_i(t) = 1 + \sum_{u=K+1}^{t} w_{u,i} + M_i(t),$$

where $(M_i(t))$ is a martingale with bounded increments. Azuma's inequality controls the martingale term, while Lemma 5 implies that posterior deviations from μ have summable tail probabilities, so the averaged targets $(1/t)\sum_{u \le t} w_u$ approach $w^\star$ fast enough. Consequently $\mathbb{E}_\mu[\tau_A] < \infty$, and therefore

$$\mathbb{E}_\mu[t_\alpha] < \infty.$$

The complete argument, including the Azuma–Hoeffding control of the martingale term and the posterior polynomial-tail estimate, is given in Appendix E.

6 From Stabilization to Correctness and Asymptotic Optimality

We now summarize how stabilization yields the final fixed-confidence and sample-complexity guarantees.

6.1 An anytime threshold for the GLR statistic

Let $C_{\exp} : \mathbb{R}_+ \to \mathbb{R}_+$ be the calibration function from Kaufmann and Koolen (2021, Theorem 7), which satisfies $C_{\exp}(x) = x + o(x)$ as $x \to \infty$. For $t \geq 1$ and $\delta \in (0, 1)$, define

$$\beta(t, \delta) := 6 \log\big(\log(t + 1) + 1\big) \; + \; 2\, C_{\exp}\left(\frac{1}{2} \log \frac{K - 1}{\delta}\right). \tag{2}$$

Proposition 2 (Anytime GLR threshold)**.** *Let β be the function defined in* (2)*. Then, for every Bernoulli parameter vector $\mu \in (0, 1)^K$ with a unique best arm and every $\delta \in (0, 1)$,*

$$\mathbb{P}_\mu\Big(\exists t \geq K + 1,\ \exists i \neq a^\star(\mu) :\ \Gamma_i(t) \geq \beta(t, \delta)\Big) \leq \delta.$$

For each wrong arm, its GLR statistic is bounded by the sum of the KL divergences from the empirical means to the true means of that arm and the true best arm, each multiplied by the corresponding number of samples. Thus a wrong arm can cross the threshold only if this sum also crosses it. Applying the time-uniform deviation inequality of Kaufmann and Koolen (2021, Theorem 7) to each such pair, as in their Section 5.1, and taking a union bound over the $K - 1$ wrong arms gives the result. The full proof is given in Appendix D.

6.2 Fixed-confidence correctness

Fix $\alpha \in (0, 1)$. Theorem 3 gives $\mathbb{E}_\mu[t_\alpha] < \infty$, hence $t_\alpha < \infty$ almost surely. After time t_α, the empirical leader is the true best arm and its GLR statistic grows at least linearly with slope $(1 - \alpha)/T^\star(\mu)$, while $\beta(t, \delta) = o(t)$. Thus the stopping time is finite almost surely. If the algorithm recommends a wrong arm, then the GLR statistic of some wrong arm must have crossed the threshold, and Proposition 2 controls this event by δ. The full proof is given in Appendix D.

6.3 Expected sample complexity

Fix $\varepsilon \in (0, 1)$ and choose $\alpha \in (0, 1/4)$ such that

$$\frac{1 + \alpha}{1 - 2\alpha} \leq 1 + \varepsilon.$$

Let t_α be the stabilization time from Theorem 3. Then $\mathbb{E}_\mu[t_\alpha] < \infty$ and, for all $t \geq t_\alpha$,

$$J_t = a^\star(\mu), \qquad \Gamma_{a^\star(\mu)}(t) \geq \frac{1 - \alpha}{T^\star(\mu)}\, t.$$

Thus it remains to compare this linear lower bound with the threshold $\beta(t, \delta)$.

Choose $\zeta \in (0, \alpha)$. Since $C_{\exp}(x) = x + o(x)$, for all sufficiently small δ,

$$2C_{\exp}\left(\frac{1}{2} \log \frac{K - 1}{\delta}\right) \leq (1 + \zeta) \log \frac{1}{\delta} + (1 + \zeta) \log(K - 1).$$

Hence, with

$$h_\zeta(t) := 6 \log(\log(t + 1) + 1) + (1 + \zeta) \log(K - 1),$$

we have

$$\beta(t, \delta) \leq (1 + \zeta) \log \frac{1}{\delta} + h_\zeta(t)$$

for all $t \geq 1$ and all sufficiently small δ. Since $h_\zeta(t) = o(t)$, choose a deterministic integer M_α such that

$$h_\zeta(t) \leq \frac{\alpha}{T^\star(\mu)}\, t \qquad \text{for all } t \geq M_\alpha.$$

Set

$$L_\delta := \frac{(1 + \zeta) T^\star(\mu)}{1 - 2\alpha} \log \frac{1}{\delta}, \qquad T_\delta := \lceil \max\{t_\alpha, M_\alpha, L_\delta\} \rceil.$$

For all sufficiently small δ, the stabilization inequality and the threshold bound give

$$\begin{aligned}\Gamma_{J_{T_\delta}}(T_\delta) = \Gamma_{a^\star(\mu)}(T_\delta) &\geq \frac{1-\alpha}{T^\star(\mu)}\, T_\delta \\ &= \frac{1-2\alpha}{T^\star(\mu)}\, T_\delta + \frac{\alpha}{T^\star(\mu)}\, T_\delta \\ &\geq (1+\zeta)\log\frac{1}{\delta} + h_\zeta(T_\delta) \geq \beta(T_\delta, \delta).\end{aligned}$$

Thus $\tau_\delta \leq T_\delta$ almost surely. Taking expectations,

$$\mathbb{E}_\mu[\tau_\delta] \leq 1 + L_\delta + \mathbb{E}_\mu\Big[\big(\max\{t_\alpha, M_\alpha\} - L_\delta\big)_+\Big].$$

The random variable $\max\{t_\alpha, M_\alpha\}$ is integrable and $L_\delta \to \infty$ as $\delta \to 0$, so dominated convergence gives

$$\mathbb{E}_\mu\Big[\big(\max\{t_\alpha, M_\alpha\} - L_\delta\big)_+\Big] = o\big(\log(1/\delta)\big).$$

Consequently,

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\log(1/\delta)} \leq \frac{1+\zeta}{1-2\alpha} T^\star(\mu) \leq (1+\varepsilon) T^\star(\mu).$$

Letting $\varepsilon \downarrow 0$ gives Theorem 2. Together with the classical lower bound of Garivier and Kaufmann (2016), this yields asymptotic optimality of Algorithm 1. The full proof is given in Appendix D.

7 Limitations

Our theoretical guarantees are established only for general K-armed Bernoulli best-arm identification in the frequentist fixed-confidence setting. In particular, this paper does not prove analogous results for broader one-parameter exponential families or for other pure-exploration tasks such as top-m identification or thresholding bandits. The fixed-confidence correctness guarantee is non-asymptotic in δ, but the optimality guarantee for expected sample complexity is asymptotic as $\delta \to 0$. Moreover, the exploration and stabilization proofs leave many instance-dependent constants implicit and therefore do not provide explicit non-asymptotic bounds on the expected sample complexity.

The deterministic reduction from a long gap to a long block (Lemma 9), the repeated application of bounds on conditional selection probabilities, and the martingale comparison between pull counts and cumulative target allocations (Appendix E) do not rely on Bernoulli rewards. The quantitative estimates used here, however, rely on the Beta–Binomial identity, bounds on the Beta density for anti-concentration, lower bounds on posterior mass above a fixed threshold, and properties of the Bernoulli KL divergence. Extending the proof would require analogous bounds on posterior probabilities at the data-dependent times used in the long-gap argument, together with suitable regularity of the oracle allocation and lower bounds on the allocation assigned to under-sampled arms. These requirements are not verified for other reward models in this paper.

Our guarantees for exploration and expected sample complexity assume the exact allocation $w^\star(\theta_t)$ at every round. We do not establish these guarantees for allocations computed with a fixed numerical tolerance; see Appendix H.2.

8 Conclusion

We studied fixed-confidence best-arm identification in Bernoulli bandits without forced exploration. We proposed a posterior-tracking algorithm that repeatedly draws a bandit instance from the posterior, computes the oracle allocation for that instance, and combines this sampling rule with a Bernoulli generalized likelihood ratio stopping rule. We showed that posterior tracking alone is sufficient to sustain exploration: every arm continues to be sampled often enough, the true best arm eventually stabilizes as the leader, and the resulting procedure is correct at the prescribed confidence level with asymptotically optimal expected sample complexity. These results establish posterior tracking as an asymptotically optimal approach to Bernoulli best-arm identification.

Several directions remain open. An important next step is to extend the analysis beyond Bernoulli rewards to broader model classes, such as general one-parameter exponential families. It would also

be interesting to investigate whether the same posterior-tracking principle can be applied to other pure-exploration tasks, such as top-m identification or thresholding bandits. A further theoretical direction is to develop a non-asymptotic analysis of posterior tracking, since the present exploration and stabilization proofs leave many instance-dependent constants implicit. A minor technical refinement would be to remove the requirement of one initial sample from each arm; this should follow from integrability of the time until every arm has been sampled at least once.

More practically, posterior-based algorithms may be attractive not only for their statistical behavior but also from a systems perspective. Thompson sampling has long been viewed as easy to implement and empirically competitive in real recommendation and advertising tasks (Chapelle and Li, 2011). Recent industry evidence likewise shows that top-two Thompson sampling can be deployed in an online-advertising platform (Katsuragawa et al., 2025). These precedents suggest that posterior tracking may also admit practical variants, although establishing and evaluating such variants is beyond the scope of the present paper.

Broader impacts. On the positive side, more sample-efficient fixed-confidence exploration may reduce the number of interactions spent on clearly inferior actions in online experimentation, thereby lowering experimental costs and unnecessary user exposure to poor choices. On the negative side, if methods of this kind are deployed in recommendation, advertising, or related decision systems, they may also accelerate optimization toward narrow engagement or revenue metrics, reinforce existing exposure biases, or amplify feedback loops in sequential decision making. This paper does not evaluate these potential societal effects or possible mitigation strategies.

Acknowledgments and Disclosure of Funding

Kaito Ariu thanks Kenshi Abe, Yuma Fujimoto, and Kodai Kureishi for helpful discussions on an earlier version of this manuscript, and OpenAI for providing access to ChatGPT through the ChatGPT for Academic Researchers program. Kaito Ariu was supported by JSPS KAKENHI Grant Number JP25K21291 and JST BOOST Program Japan Grant Number JPMJBY25A0. Junpei Komiyama was supported by the MBZUAI Start-up Fund BF0121.

References

Bandyopadhyay, A., Juneja, S., and Agrawal, S. (2024). Optimal Top-Two method for best arm identification and fluid analysis. In *Advances in Neural Information Processing Systems*, volume 37.

Barrier, A., Garivier, A., and Kocák, T. (2022). A non-asymptotic approach to Best-Arm identification for Gaussian bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 10078–10109. PMLR.

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Degenne, R. and Koolen, W. M. (2019). Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32.

Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems*, volume 32.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, volume 25, pages 3212–3220.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027. PMLR.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil' UCB : An optimal exploration algorithm for Multi-Armed bandits. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 423–439. PMLR.

Jang, K., Komiyama, J., and Yamazaki, K. (2024). Fixed confidence best arm identification in the Bayesian setting. In *Advances in Neural Information Processing Systems*, volume 37.

Jourdan, M., Degenne, R., Baudry, D., de Heide, R., and Kaufmann, E. (2022). Top Two algorithms revisited. In *Advances in Neural Information Processing Systems*, volume 35.

Juneja, S. and Krishnasamy, S. (2019). Sample complexity of partition identification using multi-armed bandits. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1824–1852. PMLR.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pages 655–662. Omnipress.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1238–1246. PMLR.

Katsuragawa, D., Kaneko, Y., Ariu, K., and Abe, K. (2025). Efficient creative selection in online advertising using Top-Two Thompson sampling. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 1090–1091.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42.

Kaufmann, E. and Kalyanakrishnan, S. (2013). Information complexity in bandit subset selection. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 228–251. PMLR.

Kaufmann, E. and Koolen, W. M. (2021). Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44.

Kaufmann, E., Koolen, W. M., and Garivier, A. (2018). Sequential test for the lowest mean: From Thompson to murphy sampling. In *Advances in Neural Information Processing Systems*, volume 31.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T., editors, *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 199–213. Springer.

Lee, J., Honda, J., and Sugiyama, M. (2024). Thompson exploration with best challenger rule in best arm identification. In *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pages 646–661. PMLR.

Liu, X., Qin, C., and You, W. (2025). Pure exploration via Frank-Wolfe self-play. *arXiv preprint arXiv:2509.19901*.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1690–1698. PMLR.

Mukherjee, S., Tripathy, A. S., and Nowak, R. (2022). Chernoff sampling for active testing and extension to active regression. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7384–7432. PMLR.

Poiani, R., Bernasconi, M., and Celli, A. (2025). Pure exploration with infinite answers. *arXiv preprint arXiv:2505.22473*.

Qin, C., Klabjan, D., and Russo, D. (2017). Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems*, volume 30, pages 5381–5391.

Qin, C. and You, W. (2025). Dual-directed algorithm design for efficient pure exploration. *Operations Research*, 74(2):1104–1125.

Russo, D. (2020). Simple Bayesian algorithms for Best-Arm identification. *Operations Research*, 68(6):1625–1647.

Shang, X., de Heide, R., Menard, P., Kaufmann, E., and Valko, M. (2020). Fixed-confidence guarantees for Bayesian Best-Arm identification. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1823–1832. PMLR.

Tabata, K., Komiyama, J., Nakamura, A., and Komatsuzaki, T. (2023). Posterior tracking algorithm for classification bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10994–11022. PMLR.

Tuynman, A. and Degenne, R. (2025). The batch complexity of bandit pure exploration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 60442–60468. PMLR.

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. (2021). Fast pure exploration via Frank-Wolfe. In *Advances in Neural Information Processing Systems*, volume 34.

Wang, S. and Zhu, J. (2022). Thompson sampling for (combinatorial) pure exploration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23470–23483. PMLR.

You, W., Qin, C., Wang, Z., and Yang, S. (2023). Information-directed selection for Top-Two algorithms. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2850–2851. PMLR.

Contents

1 Introduction **1**

1.1 Contributions 1

1.2 Related work 2

2 Setting **3**

2.1 Beta posterior and oracle quantities 3

3 Algorithm **4**

3.1 Posterior-tracking sampling rule 4

3.2 GLR statistic and stopping rule 4

4 Main Results **5**

4.1 Correctness and sample-complexity guarantees 5

4.2 From logarithmic exploration to stabilization 5

5 Sufficient Exploration under Posterior Tracking **6**

5.1 Summable logarithmic exploration 6

5.2 From accurate allocations to stabilization 7

5.3 Integrable stabilization time 7

6 From Stabilization to Correctness and Asymptotic Optimality **7**
6.1 An anytime threshold for the GLR statistic 8
6.2 Fixed-confidence correctness 8
6.3 Expected sample complexity 8

7 Limitations **9**

8 Conclusion **9**

A Additional Related Work **14**

B Table of notation for the appendix **14**

C Technical preliminaries **16**
C.1 Bernoulli KL conventions 16
C.2 Binomial KL tail 16
C.3 Posterior construction and local regularity 16
C.4 Oracle regularity near μ 17
C.5 One-arm posterior tail ingredients 18
C.5.1 Proof of Lemma 1 18

D Proofs of the fixed-confidence guarantees **19**
D.1 Proof of Proposition 2 19
D.2 Proof of Theorem 1 20
D.3 Proof of Theorem 2 20

E Proof of Theorem 3 **22**

F Proof of summable logarithmic exploration **25**
F.1 Long-gap proof of Proposition 1 26
F.1.1 Auxiliary ingredients for Lemma 10 29
F.1.2 Proof of Lemma 10 30
F.1.3 Proof of Lemma 7 43
F.1.4 Proof of Lemma 8 44
F.1.5 Proof of Lemma 11 44
F.1.6 Proof of Lemma 12 45
F.1.7 Proof of Lemma 13 47
F.1.8 Proof of Lemma 14 48

G Proof of Lemma 5 **48**

H Numerical experiments **49**
H.1 Experimental setup and stopping thresholds 49

H.2 Oracle allocation computation 50
H.3 Main results from 100 trajectories 51
H.4 Comparison of stopping thresholds 53
H.5 Comparison for different numbers of arms 54
H.6 Computation time for the oracle allocation 55

A Additional Related Work

Classical fixed-confidence BAI. The fixed-confidence BAI literature first developed through confidence-bound, gap-based, and PAC-style procedures, while Kaufmann et al. (2016) clarified the complexity viewpoint for both the fixed-budget and fixed-confidence settings. Unified gap-based exploration (UGapE) (Gabillon et al., 2012), LUCB (Kalyanakrishnan et al., 2012), and the KL-based subset-selection analysis of Kaufmann and Kalyanakrishnan (2013), together with later instance-sensitive algorithms such as Karnin et al. (2013) and Jamieson et al. (2014), established finite-sample and instance-dependent guarantees in classical stochastic bandit models.

Lower-bound tracking and game formulations. A parallel line pursued the lower-bound viewpoint more directly through oracle allocations and related optimization formulations. Garivier and Kaufmann (2016) characterized the fixed-confidence complexity and introduced Track-and-Stop, while Degenne et al. (2019) recast pure exploration as an unknown game and Wang et al. (2021) developed Frank-Wolfe-based sampling to approximate lower-bound allocations efficiently. Barrier et al. (2022) study non-asymptotic best-arm identification for Gaussian bandits with an exploration-biased sampling rule. A recent adjacent game-theoretic line is Liu et al. (2025), who reformulate pure exploration as a concave-convex saddle-point problem with a mixed-strategy skeptic and propose Frank-Wolfe Self-Play, proving game-value convergence for best-arm identification in linear bandits. They also introduce a posterior-sampling-based learning variant, but its theoretical analysis is left for future work. Relatedly, Degenne and Koolen (2019) extend the lower-bound and tracking perspective to pure-exploration problems with multiple correct answers, where continuity and convexity issues alter Track-and-Stop-style optimality. These methods share the lower-bound perspective, but the FWSP results concern game-value convergence rather than a Bernoulli fixed-confidence stopping theorem.

Thompson-style pure exploration beyond ordinary BAI. TS-Explore (Wang and Zhu, 2022) uses independent Thompson-style samples for combinatorial pure exploration. Its main goal is to reduce the conservatism of summed armwise confidence bounds in combinatorial settings, and in the ordinary multi-armed case it obtains an asymptotically optimal upper bound in the classical gap-complexity scale $H_m \log(1/\delta)$. This notion of optimality differs from the characteristic-time optimality of Garivier and Kaufmann (2016) considered in this paper, which requires matching the exact Bernoulli fixed-confidence lower-bound constant $T^\star(\mu)$ in expected stopping time. TS-Explore also does not track posterior-induced oracle allocations.

Related pure-exploration settings. Several related settings differ from the ordinary Bernoulli fixed-confidence problem considered in the main text. Locatelli et al. (2016) study thresholding bandits under a fixed-budget objective rather than a fixed-confidence objective, and Jang et al. (2024) analyze fixed-confidence BAI in the Bayesian setting, where the bandit instance itself is drawn from a prior. More broadly, Juneja and Krishnasamy (2019) formulate partition identification as a general pure-exploration framework that subsumes several query types beyond ordinary best-arm identification. Recent developments on active testing, batched pure exploration, and infinite-answer pure exploration include Mukherjee et al. (2022), Tuynman and Degenne (2025), and Poiani et al. (2025).

B Table of notation for the appendix

The following table summarizes the main notation used in the appendix, ordered by first appearance.

Symbol	Meaning
Logarithmic exploration preliminaries	
$\Delta_{\min}$	Best-arm gap $\mu_{a^\star(\mu)} - \max_{j \neq a^\star(\mu)} \mu_j$.
r_μ	Positive fixed local radius around μ, chosen so that $\overline{B}_\infty(\mu, r_\mu) \subset (0,1)^K$, $r_\mu \leq \Delta_{\min}/8$, and the oracle-continuity implication (3) holds; used in the empirical good event, posterior concentration bounds, and the definition of Good_s.
$\mathcal{E}_t$	Good empirical event on which every empirical mean lies within $r_\mu/2$ of the corresponding true mean at time t.
C_u	Geometric-series constant in the stopping-time Hoeffding bound with radius u.
$C_{r_\mu/2}$	Specialization of C_u at $u = r_\mu/2$.
Stopping and sample complexity	
$C^\star(x)$	Appendix shorthand for $\max_{w \in \Delta_K} F(x,w) = F(x, w^\star(x))$.
$T^\star(\mu)$	Information-theoretic characteristic time, defined here as $1/C^\star(\mu)$.
$\beta(t,\delta)$	Time-uniform GLR threshold used by the stopping rule.
τ_δ	Stopping time $\inf\{t \geq K+1 : \Gamma_{J_t}(t) \geq \beta(t,\delta)\}$.
$\hat{a}$	Recommendation returned at the stopping time, namely J_{τ_δ}.
Long-gap notation	
$\tau_{i,j}$	Time of the j-th play of arm i, with the convention $\tau_{i,j} = t+1$ if arm i has been played fewer than j times by horizon t.
L	Length threshold forcing the existence of a long i-gap when $N_i(t) < A \log t$.
$E_{i,j}$	Event that the gap between the j-th and $(j+1)$-st plays of arm i has length at least L.
$\mu^-_{a^\star(\mu)}$	Shifted best-arm threshold $\mu_{a^\star(\mu)} - \Delta_{\min}/4$ used in the best-arm posterior tail bound.
R_μ	KL ratio comparing the best-arm KL term to the threshold KL term below $\mu^-_{a^\star(\mu)}$.
$\mathrm{kl}_-(x,u)$	Truncated KL term: $\mathrm{kl}(x,u)$ for $x \leq u$ and 0 for $x > u$.
$\mathcal{S}_s$	Saturated set at time s, i.e. the arms with at least $A \log t$ samples.
I	Gap interval $\{\tau_{i,j}+1, \ldots, \tau_{i,j+1}-1\}$ associated with a fixed long i-gap.
M	Deterministic upper bound on the number of interruptions inside I.
L'	Guaranteed length of a long non-interruption subblock after removing interruptions from I.
T_q	Time just before the q-th candidate non-interruption block inside the gap interval I.
B_q	q-th non-interruption block cut out from I.
$\widetilde{B}_q$	First L' rounds of the block B_q.
$\widetilde{E}_{i,j,q}$	Event that the block B_q has length at least L'.
δ_t	Separation scale $(\log t)^{-4}$ used in the anti-concentration and oracle-mass estimates.
$\mathcal{U}$	Unsaturated set at time T_q, namely $\{1,\ldots,K\} \setminus \mathcal{S}_{T_q}$.
Ξ_ℓ	Event that the next ℓ rounds after T_q sample only arms in $\mathcal{S}_{T_q}$.
$\mathcal{H}_{q,\ell}$	Conditioning sigma-field from Lemma 11, used for the one-step analysis at round $T_q + \ell$.
Good_s	Event that $\lvert\theta_{k,s} - \mu_k\rvert \leq r_\mu$ for every $k \in \mathcal{S}_{T_q}$.
Bad	Number of rounds in $\widetilde{B}_q$ on which Good_s fails.
Φ_q	Conditional posterior mass that the true best-arm draw at time T_q+1 lies above $\mu_{a^\star(\mu)} - \Delta_{\min}/4$.
Final stabilization proof	
$\mathcal{N}_t$	Event that every arm already has at least $A \log t$ samples at time $t-1$.
r_0	Auxiliary radius in $(0, r_\mu]$ used in the polynomial posterior-tail lemma and the final stabilization proof.
τ_A	Allocation stabilization time after which $N(s)/s$ stays uniformly close to $w^\star(\mu)$.
τ_B	Time immediately after the last empirical failure while the allocation ratios are close to $w^\star(\mu)$; $K+1$ if no such failure occurs.
w_t	Random target allocation vector $w^\star(\theta_t)$ used by posterior-tracking at round t.
$M_i(t)$	Martingale comparing the realized pull count of arm i with the cumulative target weights up to time t.

C Technical preliminaries

C.1 Bernoulli KL conventions

For $q \in (0,1)$, we use the standard boundary conventions

$$\mathrm{kl}(0,q) = \log \frac{1}{1-q}, \qquad \mathrm{kl}(1,q) = \log \frac{1}{q},$$

with $\mathrm{kl}(0,0) = \mathrm{kl}(1,1) = 0$, $\mathrm{kl}(p,0) = +\infty$ for $p \in (0,1]$, and $\mathrm{kl}(p,1) = +\infty$ for $p \in [0,1)$.

C.2 Binomial KL tail

Lemma 2 (Binomial KL tail bound (Chernoff)). *Let $Y \sim \mathrm{Bin}(N,p)$ with $N \in \mathbb{N}$ and $p \in (0,1)$. Then for any $m \in \{0,1,\dots,N\}$,*

$$\mathbb{P}(Y \geq m) \leq \exp\Big(-N\,\mathrm{kl}\big(m/N,\,p\big)\Big) \quad \textit{if } m/N > p,$$

and

$$\mathbb{P}(Y \leq m) \leq \exp\Big(-N\,\mathrm{kl}\big(m/N,\,p\big)\Big) \quad \textit{if } m/N < p.$$

Proof. For the upper tail, let $\lambda > 0$ and use Markov's inequality:

$$\mathbb{P}(Y \geq m) = \mathbb{P}(e^{\lambda Y} \geq e^{\lambda m}) \leq e^{-\lambda m}\,\mathbb{E}[e^{\lambda Y}] = \exp\Big(-\lambda m + N\log(1-p+pe^{\lambda})\Big).$$

Minimizing the exponent over $\lambda > 0$ yields the bound $\mathbb{P}(Y \geq m) \leq \exp\big(-N\mathrm{kl}(m/N,p)\big)$ when $m/N > p$. The lower tail follows similarly by applying the upper-tail bound to $N - Y \sim \mathrm{Bin}(N, 1-p)$, or by the same Chernoff argument with $\lambda < 0$. □

C.3 Posterior construction and local regularity

Assume an independent Beta prior:

$$\mu_i \sim \mathrm{Beta}(\alpha_i, \beta_i), \qquad \alpha_i, \beta_i \in \mathbb{N},\ \alpha_i \geq 1,\ \beta_i \geq 1,$$

independent across i. Then the posterior factorizes and for each arm i,

$$\mu_i \mid \mathcal{F}_t \sim \mathrm{Beta}(\alpha_i + S_i(t),\ \beta_i + N_i(t) - S_i(t)).$$

We denote the posterior at time t by Π_t and a posterior draw by $\theta_{t+1} \sim \Pi_t$ (so $\theta_t \sim \Pi_{t-1}$). Equivalently, after enlarging the probability space if needed, for each round $t > K$ and each arm i there exists an auxiliary variable

$$U_{t,i} \sim \mathrm{Unif}(0,1),$$

such that the full array $(U_{t,i})_{t>K,\ 1\leq i\leq K}$ is i.i.d. and independent of the original bandit probability space. Each $\theta_{t,i}$ is a measurable function of $\mathcal{F}_{t-1}$ and $U_{t,i}$, so conditionally on $\mathcal{F}_{t-1}$ the variables $(\theta_{t,i})_{i=1}^K$ are independent. Moreover, for each deterministic time $t > K$,

$$(U_{t,i})_{i=1}^K \perp \sigma\big(\mathcal{F}_{t-1}, \theta_{K+1}, \dots, \theta_{t-1}\big).$$

Indeed, each past posterior draw θ_m is a measurable function of $\mathcal{F}_{m-1}$ and the contemporaneous innovation vector $(U_{m,i})_{i=1}^K$, so the enlarged past above is contained in the sigma-field generated by the original bandit process together with the innovation vectors from rounds strictly before t, which is independent of the current vector $(U_{t,i})_{i=1}^K$. This is the deterministic-time independence used later to freeze posterior parameters while keeping current unsaturated draws fresh.

Positivity of $w^\star(x)$. Let $x \in (0,1)^K$ have a unique best arm $a = a^\star(x)$ and define the gap $\Delta(x) := x_a - \max_{j\neq a} x_j > 0$. Write

$$C^\star(x) := \max_{w\in\Delta_K} F(x,w) = F(x, w^\star(x)).$$

If a weight vector $w \in \Delta_K$ has a zero entry, then $F(x,w) = 0$: if $w_i = 0$ for some $i \neq a$, take $\lambda \in (0,1)^K$ with $\lambda_i = x_a$ and $\lambda_k = x_k$ for $k \neq i$; if $w_a = 0$, pick any $j \neq a$ and take

$\lambda_a = x_j$ and $\lambda_k = x_k$ for $k \neq a$. In both cases $\lambda \in \mathrm{Alt}(x)$ and $\sum_{k=1}^K w_k d(x_k, \lambda_k) = 0$. On the other hand, for $\bar{w} := (1/K, \dots, 1/K)$, any $\lambda \in \mathrm{Alt}(x)$ admits $j \neq a$ with $\lambda_j \geq \lambda_a$, hence either $|\lambda_a - x_a| \geq \Delta(x)/2$ or $|\lambda_j - x_j| \geq \Delta(x)/2$. By Pinsker's inequality $d(p,q) \geq 2(p-q)^2$, this yields

$$F(x, \bar{w}) \geq \frac{\Delta(x)^2}{2K} > 0,$$

so $C^\star(x) > 0$ and therefore any maximizer $w^\star(x) \in \arg\max_w F(x,w)$ must satisfy $w_i^\star(x) > 0$ for all i.

C.4 Oracle regularity near μ

Let

$$\Delta_{\min} := \mu_{a^\star(\mu)} - \max_{j \neq a^\star(\mu)} \mu_j > 0.$$

In the Bernoulli case, the local regularity needed below follows from the oracle-allocation results of Garivier and Kaufmann (2016); we record it here for convenience.

Proposition 3 (Oracle regularity near μ in the Bernoulli case)**.** *Assume that $\mu \in (0,1)^K$ has a unique best arm. Then there exists a neighborhood U of μ such that: (i) $a^\star(\theta) = a^\star(\mu)$ for all $\theta \in U$; (ii) $w^\star(\theta)$ is unique for all $\theta \in U$ and $\theta \mapsto w^\star(\theta)$ is continuous on U.*

Proof. Let $a := a^\star(\mu)$ and define

$$\Delta_\mu := \mu_a - \max_{j \neq a} \mu_j > 0.$$

Choose

$$\rho := \min\Big\{ \frac{\Delta_\mu}{3},\ \frac{1}{2} \min_{1 \leq i \leq K} \{\mu_i, 1-\mu_i\} \Big\} > 0, \qquad U := B_\infty(\mu, \rho).$$

Then $U \subset (0,1)^K$. Moreover, for every $\theta \in U$ and every $j \neq a$,

$$\theta_a - \theta_j \geq (\mu_a - \mu_j) - |\theta_a - \mu_a| - |\theta_j - \mu_j| \geq \Delta_\mu - 2\|\theta - \mu\|_\infty > \Delta_\mu - 2\rho > 0.$$

Hence $a^\star(\theta) = a$ for all $\theta \in U$, proving (i).

For (ii), define

$$\mathcal{S}_a := \{x \in (0,1)^K :\ x_a > x_j\ \ \forall j \neq a\}.$$

By (i), $U \subset \mathcal{S}_a$. For $x \in \mathcal{S}_a$ and $j \neq a$, define

$$g_{a,j}^x(u) := (1+u)\, I_{1/(1+u)}(x_a, x_j), \qquad u \geq 0,$$

where

$$I_\alpha(p,q) := \alpha\, \mathrm{kl}\big(p, \alpha p + (1-\alpha)q\big) + (1-\alpha)\, \mathrm{kl}\big(q, \alpha p + (1-\alpha)q\big).$$

Since $x_a > x_j$, each $g_{a,j}^x$ is a strictly increasing bijection from $[0,\infty)$ onto $[0, \mathrm{kl}(x_a, x_j))$; let $x_{a,j}^x : [0, \mathrm{kl}(x_a, x_j)) \to [0,\infty)$ denote its inverse. Garivier and Kaufmann (2016, Theorem 5), written with arm a as the distinguished best arm, show that there exists a unique scalar $y^\star(x)$ such that

$$w_a^\star(x) = \frac{1}{1 + \sum_{j \neq a} x_{a,j}^x(y^\star(x))}, \qquad w_j^\star(x) = \frac{x_{a,j}^x(y^\star(x))}{1 + \sum_{k \neq a} x_{a,k}^x(y^\star(x))} \quad (j \neq a).$$

In particular, $w^\star(x)$ is uniquely determined for every $x \in \mathcal{S}_a$. Moreover, Garivier and Kaufmann (2016, Proposition 6(2)) show that $x \mapsto w^\star(x)$ is continuous at every bandit instance with a unique best arm. In particular, it is continuous on $\mathcal{S}_a$, hence on U. This proves (ii). □

By Proposition 3, continuity of $w^\star$ at μ, and positivity of $w^\star(\mu)$, we fix a radius $r_\mu > 0$ such that

$$\overline{B}_\infty(\mu, r_\mu) \subset (0,1)^K, \qquad r_\mu \leq \frac{\Delta_{\min}}{8},$$

and such that the following also holds for all θ:

$$\|\theta - \mu\|_\infty \leq r_\mu \quad \Longrightarrow \quad \|w^\star(\theta) - w^\star(\mu)\|_\infty \leq \frac{1}{4} w_{\min}, \qquad w_{\min} := \min_i w_i^\star(\mu) > 0. \tag{3}$$

Hence on $\{\|\theta - \mu\|_\infty \leq r_\mu\}$ one has $w_i^\star(\theta) \geq \frac{3}{4} w_{\min}$ for all i.

C.5 One-arm posterior tail ingredients

We first prove a one-arm posterior-sample concentration inequality. The next two lemmas isolate the standard chain used repeatedly later: first control one Beta posterior draw conditional on its sufficient statistics, and then use that estimate directly in the stopping-time and time-polynomial arguments below.

Lemma 3 (Conditional tail for a Beta posterior draw)**.** *Fix $p \in (0,1)$ and $u \in \big(0, \min\{p, 1-p\}\big)$. Let $\alpha, \beta \in \mathbb{N}$ with $\alpha \geq 1$ and $\beta \geq 1$. There exist constants $c_0 = c_0(p,u) > 0$ and $C_0 = C_0(p,u,\alpha,\beta) > 0$ such that for all $n \geq 1$ and all $s \in \{0,1,\dots,n\}$ satisfying $|s/n - p| \leq u/2$, if $\theta \sim \mathrm{Beta}(\alpha + s, \beta + n - s)$ then*

$$\mathbb{P}\big(|\theta - p| > u\big) \leq C_0 e^{-c_0 n}.$$

Proof. Let $N := n + \alpha + \beta - 1$. Fix s with $|s/n - p| \leq u/2$ and write $\hat{p} := s/n$. By the Beta–Binomial identity,

$$\mathbb{P}(\theta \geq x) = \mathbb{P}\big(Y \leq \alpha + s - 1\big), \qquad Y \sim \mathrm{Bin}(N, x),$$

and

$$\mathbb{P}(\theta \leq x) = \mathbb{P}\big(Y \geq \alpha + s\big), \qquad Y \sim \mathrm{Bin}(N, x).$$

Upper tail. Let $x := p + u$. Since $\hat{p} \leq p + u/2$, we have $\alpha + s - 1 \leq \alpha - 1 + n(p + u/2)$. For n large enough (depending on p, u, α, β), this implies $(\alpha - 1 + n(p + u/2))/N \leq p + 3u/4 < p + u$. Since Y is integer, $\{Y \leq \alpha + s - 1\} \subseteq \{Y \leq \lfloor N(p + 3u/4) \rfloor\}$ and therefore

$$\mathbb{P}(\theta \geq p + u) \leq \mathbb{P}\big(Y \leq \lfloor N(p + 3u/4) \rfloor\big) \leq \exp\Big(-N\,\mathrm{kl}\big(p + 3u/4,\, p + u\big)\Big),$$

by Lemma 2.

Lower tail. Let $x := p - u$. Since $\hat{p} \geq p - u/2$, we have $\alpha + s \geq \alpha + n(p - u/2)$. For n large enough, again depending on p, u, α, β, this implies $(\alpha + n(p - u/2))/N \geq p - 3u/4 > p - u$ and hence $\{Y \geq \alpha + s\} \subseteq \{Y \geq \lceil N(p - 3u/4) \rceil\}$. By Lemma 2,

$$\mathbb{P}(\theta \leq p - u) \leq \mathbb{P}\big(Y \geq \lceil N(p - 3u/4) \rceil\big) \leq \exp\Big(-N\,\mathrm{kl}\big(p - 3u/4,\, p - u\big)\Big).$$

Set

$$c_0 := \min\{\mathrm{kl}(p + 3u/4, p + u),\ \mathrm{kl}(p - 3u/4, p - u)\} > 0,$$

which depends only on p, u. Let $n_0 = n_0(p, u, \alpha, \beta)$ be large enough that the two preceding inclusions hold for all $n \geq n_0$. For such n,

$$\mathbb{P}(|\theta - p| > u) \leq 2e^{-c_0 N} \leq 2e^{-c_0 n}.$$

Choosing $C_0 = C_0(p, u, \alpha, \beta) \geq 2$ large enough absorbs the finitely many cases $1 \leq n < n_0$ and the finitely many admissible values of s. □

C.5.1 Proof of Lemma 1

Proof of Lemma 1. If $\hat{\mu}_i(t) < \max_{k \neq i} \hat{\mu}_k(t)$, take convex combinations of $\hat{\mu}(t)$ with $(1/2, \dots, 1/2)$, giving the latter vector a positive weight that tends to zero. These vectors lie in $(0,1)^K$ and preserve all coordinate orderings, so they belong to Alt_i. Each KL term tends to zero, including when an empirical mean equals 0 or 1. Since KL divergence is nonnegative, this proves $\Gamma_i(t) = 0$.

Assume now that $\hat{\mu}_i(t) = \max_k \hat{\mu}_k(t)$. Since $\mathrm{Alt}_i = \bigcup_{j \neq i} \{\lambda : \lambda_j \geq \lambda_i\}$, we first choose which challenger j catches up with arm i, and only then solve the remaining two-variable constrained problem:

$$\Gamma_i(t) = \min_{j \neq i} \inf_{\lambda : \lambda_j \geq \lambda_i} \sum_{k=1}^{K} N_k(t)\,\mathrm{kl}(\hat{\mu}_k(t), \lambda_k).$$

For fixed j, each coordinate $k \notin \{i, j\}$ has infimum KL contribution zero, approached by letting λ_k tend to $\hat{\mu}_k(t)$ from within $(0,1)$. These coordinates do not affect the constraint $\lambda_j \geq \lambda_i$, so it remains to solve

$$\inf_{\lambda_j \geq \lambda_i} N_i(t)\mathrm{kl}(\hat{\mu}_i(t), \lambda_i) + N_j(t)\mathrm{kl}(\hat{\mu}_j(t), \lambda_j).$$

If $\hat{\mu}_i(t) > \hat{\mu}_j(t)$, the line segment from any feasible pair with $\lambda_j > \lambda_i$ toward $(\hat{\mu}_i(t), \hat{\mu}_j(t))$ reaches the diagonal within $(0,1)^2$, and the objective decreases by convexity. If $\hat{\mu}_i(t) = \hat{\mu}_j(t)$, the infimum is zero and is approached by taking both coordinates equal and tending to their common empirical mean from within $(0,1)$. Thus in either case the infimum can be computed on the diagonal. Set $\lambda_i = \lambda_j = u$ and minimize $u \mapsto N_i(t)\mathrm{kl}(\hat{\mu}_i(t), u) + N_j(t)\mathrm{kl}(\hat{\mu}_j(t), u)$ over $u \in (0,1)$. This function is strictly convex and its derivative is

$$\frac{N_i(t)(u - \hat{\mu}_i(t)) + N_j(t)(u - \hat{\mu}_j(t))}{u(1-u)},$$

which vanishes at $u = \hat{m}_{i,j}(t)$ when $\hat{m}_{i,j}(t) \in (0,1)$. If $\hat{m}_{i,j}(t) \in \{0,1\}$, then the same value is obtained by taking the limit $u \to \hat{m}_{i,j}(t)$ within $u \in (0,1)$; by the boundary convention for kl, this limit is exactly $Z_{i,j}(t)$. Thus in all cases the infimum equals $Z_{i,j}(t)$. Taking the minimum over $j \neq i$ proves the lemma. □

D Proofs of the fixed-confidence guarantees

D.1 Proof of Proposition 2

Proof. Fix the true best arm $a^\star(\mu)$, and for each wrong arm $i \neq a^\star(\mu)$ define

$$E_i := \Big\{ \exists t \geq K+1 : \ \Gamma_i(t) \geq \beta(t,\delta) \Big\}.$$

For every $t \geq K+1$, the definition of $\Gamma_i(t)$ gives

$$\begin{aligned}
\Gamma_i(t) &= \inf_{\lambda \in \mathrm{Alt}_i} \sum_{k=1}^{K} N_k(t)\mathrm{kl}(\hat{\mu}_k(t), \lambda_k) \\
&\leq \inf_{\substack{\lambda \in (0,1)^K \\ \lambda_{a^\star(\mu)} \geq \lambda_i}} \sum_{k=1}^{K} N_k(t)\mathrm{kl}(\hat{\mu}_k(t), \lambda_k) \\
&= \inf_{\substack{\lambda_i, \lambda_{a^\star(\mu)} \in (0,1) \\ \lambda_{a^\star(\mu)} \geq \lambda_i}} \sum_{j \in \{i, a^\star(\mu)\}} N_j(t)\mathrm{kl}(\hat{\mu}_j(t), \lambda_j) \\
&\leq \sum_{j \in \{i, a^\star(\mu)\}} N_j(t)\mathrm{kl}(\hat{\mu}_j(t), \mu_j).
\end{aligned}$$

The first inequality restricts the alternatives to those satisfying $\lambda_{a^\star(\mu)} \geq \lambda_i$. For the second equality, every coordinate outside $\{i, a^\star(\mu)\}$ can approach its empirical mean from within $(0,1)$, so its KL contribution tends to zero, including when the empirical mean equals 0 or 1. The last inequality follows by taking $\lambda_i = \mu_i$ and $\lambda_{a^\star(\mu)} = \mu_{a^\star(\mu)}$, which is feasible because $\mu_{a^\star(\mu)} > \mu_i$. These steps do not require any particular empirical ordering of the arms.

We apply Kaufmann and Koolen (2021, Theorem 7) to the fixed pair $\{i, a^\star(\mu)\}$, following the best-arm identification argument in their Section 5.1. The theorem applies to adaptive sampling rules, including randomized ones. Since initialization ensures $N_j(t) \geq 1$, applying the theorem to these two arms with deviation parameter $2x$ gives, for every $x > 0$,

$$\mathbb{P}_\mu \left(\begin{array}{c} \exists t \geq K+1 : \\ \displaystyle\sum_{j \in \{i, a^\star(\mu)\}} \max\big\{ N_j(t)\mathrm{kl}(\hat{\mu}_j(t), \mu_j) - 3\log\big(1 + \log N_j(t)\big),\, 0 \big\} \geq 2C_{\exp}(x) \end{array} \right) \leq e^{-2x}.$$

Each maximum takes the positive part of the expression inside it. Replacing a term by its positive part can only increase the sum, so the same probability bound holds with the maxima removed. Now take

$$x = \frac{1}{2}\log\frac{K-1}{\delta}.$$

For every $t \geq K+1$, the bounds $1 \leq N_j(t) \leq t$ and the definition of $\beta(t,\delta)$ imply

$$\begin{aligned}
\beta(t,\delta) &= 6\log\big(\log(t+1) + 1\big) + 2C_{\exp}(x) \\
&\geq 3 \sum_{j \in \{i, a^\star(\mu)\}} \log\big(1 + \log N_j(t)\big) + 2C_{\exp}(x).
\end{aligned}$$

Combining these bounds gives

$$\begin{aligned}\mathbb{P}_\mu(E_i) &\le \mathbb{P}_\mu\left(\exists t \ge K+1 : \sum_{j\in\{i,a^\star(\mu)\}} N_j(t)\mathrm{kl}(\hat{\mu}_j(t),\mu_j) \ge \beta(t,\delta)\right)\\ &\le \mathbb{P}_\mu\left(\begin{array}{c}\exists t \ge K+1 :\\ \sum_{j\in\{i,a^\star(\mu)\}} \big[N_j(t)\mathrm{kl}(\hat{\mu}_j(t),\mu_j) - 3\log\big(1+\log N_j(t)\big)\big] \ge 2C_{\exp}(x)\end{array}\right)\\ &\le e^{-2x} = \frac{\delta}{K-1}.\end{aligned}$$

Finally, a union bound over the $K-1$ wrong arms yields

$$\begin{aligned}&\mathbb{P}_\mu(\exists t \ge K+1,\ \exists i \ne a^\star(\mu) :\ \Gamma_i(t) \ge \beta(t,\delta))\\ &\quad= \mathbb{P}_\mu\left(\bigcup_{i\ne a^\star(\mu)} E_i\right) \le \sum_{i\ne a^\star(\mu)} \mathbb{P}_\mu(E_i) \le (K-1)\frac{\delta}{K-1} = \delta.\end{aligned}$$

□

D.2 Proof of Theorem 1

Proof. Fix $\alpha \in (0,1)$. Let

$$t_\alpha := \inf\Big\{t \ge K+1 :\ J_s = a^\star(\mu) \text{ and } \Gamma_{a^\star(\mu)}(s) \ge \frac{1-\alpha}{T^\star(\mu)}\, s \text{ for all } s \ge t\Big\}.$$

Theorem 3 gives $\mathbb{E}_\mu[t_\alpha] < \infty$, hence $t_\alpha < \infty$ almost surely, and

$$\beta(t,\delta) = 6\log\big(\log(t+1)+1\big) + 2\,C_{\exp}\left(\frac{1}{2}\log\frac{K-1}{\delta}\right) = o(t) \qquad (t\to\infty).$$

Hence there exists a deterministic integer M_δ such that

$$\beta(t,\delta) \le \frac{1-\alpha}{T^\star(\mu)}\, t \qquad \text{for all } t \ge M_\delta.$$

On the event $\{t_\alpha < \infty\}$, every $t \ge \max\{t_\alpha, M_\delta\}$ then satisfies

$$J_t = a^\star(\mu) \qquad \text{and} \qquad \Gamma_{a^\star(\mu)}(t) \ge \frac{1-\alpha}{T^\star(\mu)}\, t \ge \beta(t,\delta),$$

so $\tau_\delta \le \max\{t_\alpha, M_\delta\} < \infty$ almost surely. If $\hat{a} \ne a^\star(\mu)$, then at time τ_δ the recommended arm J_{τ_δ} is wrong and satisfies

$$\Gamma_{J_{\tau_\delta}}(\tau_\delta) \ge \beta(\tau_\delta,\delta).$$

Therefore

$$\{\hat{a} \ne a^\star(\mu)\} \subseteq \Big\{\exists t \ge K+1,\ \exists i \ne a^\star(\mu) :\ \Gamma_i(t) \ge \beta(t,\delta)\Big\},$$

and Proposition 2 gives

$$\mathbb{P}_\mu(\hat{a} \ne a^\star(\mu)) \le \delta.$$

□

D.3 Proof of Theorem 2

Proof. Fix $\varepsilon \in (0,1)$ and choose $\alpha \in (0, 1/4)$ such that

$$\frac{1+\alpha}{1-2\alpha} \le 1+\varepsilon.$$

Let t_α be the stabilization time from Theorem 3 for this value of α:

$$t_\alpha := \inf\Big\{t \ge K+1 :\ J_s = a^\star(\mu) \text{ and } \Gamma_{a^\star(\mu)}(s) \ge \frac{1-\alpha}{T^\star(\mu)}\, s \text{ for all } s \ge t\Big\}.$$

Theorem 3 gives $\mathbb{E}_\mu[t_\alpha] < \infty$, hence $t_\alpha < \infty$ almost surely.

By $C_{\exp}(x) = x + o(x)$ as $x \to \infty$, we may choose $\zeta \in (0, \alpha)$ and $x_\zeta > 0$ such that

$$C_{\exp}(x) \le (1+\zeta)\, x \qquad \text{for all } x \ge x_\zeta.$$

Then there exists $\delta_\zeta \in (0,1)$ such that, for every $\delta \in (0, \delta_\zeta)$,

$$\frac{1}{2}\log\frac{K-1}{\delta} \ge x_\zeta$$

and

$$2\, C_{\exp}\left(\frac{1}{2}\log\frac{K-1}{\delta}\right) \le (1+\zeta)\log\frac{K-1}{\delta}.$$

Therefore, for every $\delta \in (0, \delta_\zeta)$ and every $t \ge 1$,

$$\beta(t,\delta) \le 6\log\big(\log(t+1)+1\big) + (1+\zeta)\log\frac{1}{\delta} + (1+\zeta)\log(K-1).$$

The term $6\log\big(\log(t+1)+1\big) + (1+\zeta)\log(K-1)$ is $o(t)$ as $t \to \infty$ and does not depend on δ. Thus we may choose a deterministic integer M_α, independent of δ, such that

$$6\log\big(\log(t+1)+1\big) + (1+\zeta)\log(K-1) \le \frac{\alpha}{T^\star(\mu)}\, t \qquad \text{for all } t \ge M_\alpha.$$

Define

$$L_\delta := \frac{(1+\zeta)T^\star(\mu)}{1-2\alpha}\log\frac{1}{\delta}.$$

Define the integer-valued random cutoff

$$T_\delta := \lceil \max\{t_\alpha, M_\alpha, L_\delta\} \rceil .$$

On the event $\{t_\alpha < \infty\}$ and for every $\delta \in (0, \delta_\zeta)$,

$$J_{T_\delta} = a^\star(\mu) \quad \text{and} \quad \Gamma_{a^\star(\mu)}(T_\delta) \ge \frac{1-\alpha}{T^\star(\mu)}\, T_\delta.$$

By construction, T_δ is large enough to ensure stabilization and to control the two threshold terms. Therefore,

$$\Gamma_{J_{T_\delta}}(T_\delta) = \Gamma_{a^\star(\mu)}(T_\delta) \ge \frac{1-\alpha}{T^\star(\mu)}\, T_\delta \;=\; \frac{1-2\alpha}{T^\star(\mu)}\, T_\delta + \frac{\alpha}{T^\star(\mu)}\, T_\delta \ge (1+\zeta)\log\frac{1}{\delta} + \frac{\alpha}{T^\star(\mu)}\, T_\delta,$$

where the last inequality uses $T_\delta \ge L_\delta$. Since $T_\delta \ge M_\alpha$,

$$6\log\big(\log(T_\delta+1)+1\big) + (1+\zeta)\log(K-1) \le \frac{\alpha}{T^\star(\mu)}\, T_\delta.$$

Combining the last two bounds, we obtain

$$\Gamma_{J_{T_\delta}}(T_\delta) \ge \beta(T_\delta, \delta).$$

Therefore, for every $\delta \in (0, \delta_\zeta)$, we have $\tau_\delta \le T_\delta$ almost surely. Taking expectations yields

$$\mathbb{E}_\mu[\tau_\delta] \le \mathbb{E}_\mu[T_\delta] \le 1 + \mathbb{E}_\mu\big[\max\{t_\alpha, M_\alpha, L_\delta\}\big] = 1 + L_\delta + \mathbb{E}_\mu\Big[\big(\max\{t_\alpha, M_\alpha\} - L_\delta\big)_+\Big].$$

Since $\max\{t_\alpha, M_\alpha\}$ is integrable and $L_\delta \to \infty$ as $\delta \to 0$, dominated convergence gives

$$\mathbb{E}_\mu\Big[\big(\max\{t_\alpha, M_\alpha\} - L_\delta\big)_+\Big] = o\big(\log(1/\delta)\big).$$

Consequently,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau_\delta]}{\log(1/\delta)} \le \frac{1+\zeta}{1-2\alpha}\, T^\star(\mu) \le (1+\varepsilon) T^\star(\mu).$$

Letting $\varepsilon \downarrow 0$ completes the proof. □

E Proof of Theorem 3

Lemma 4 (Local deterministic stabilization criterion)**.** *Assume that $\mu \in (0,1)^K$ has a unique best arm, and fix $\varepsilon \in (0,1)$. Write $a^\star := a^\star(\mu)$ and $w^\star := w^\star(\mu)$. Then there exist $\rho > 0$ and $\eta \in (0, \frac{1}{2}\min_i w_i^\star]$ such that, for every $t \geq K$, if*

$$\|\hat{\mu}(t) - \mu\|_\infty \leq \rho \quad \textit{and} \quad \|N(t)/t - w^\star\|_\infty \leq \eta,$$

then

$$J_t = a^\star \quad \textit{and} \quad \Gamma_{a^\star}(t) \geq \frac{1-\varepsilon}{T^\star(\mu)}\, t. \tag{4}$$

Proof. Let $\Delta_{\min} := \min_{j \neq a^\star}(\mu_{a^\star} - \mu_j) > 0$ and $w_{\min} := \min_{1 \leq i \leq K} w_i^\star > 0$. Since $a^\star$ is unique, there exists a neighborhood U of μ such that $a^\star(x) = a^\star$ for all $x \in U$, hence $\mathrm{Alt}(x) = \mathrm{Alt}(\mu)$ on U. On $U \times \Delta_K$, the map $(x,w) \mapsto F(x,w)$ is continuous. Indeed,

$$F(x,w) = \min_{j \neq a^\star} \inf_{\lambda_j \geq \lambda_{a^\star}} \sum_{k=1}^K w_k \,\mathrm{kl}(x_k, \lambda_k),$$

and for each fixed $j \neq a^\star$, writing $a := a^\star$, the inner infimum equals

$$g_j(x,w) := \begin{cases} w_a \,\mathrm{kl}(x_a, m_{a,j}(x,w)) + w_j \,\mathrm{kl}(x_j, m_{a,j}(x,w)), & w_a + w_j > 0, \\ 0, & w_a = w_j = 0, \end{cases}$$

where $m_{a,j}(x,w) := (w_a x_a + w_j x_j)/(w_a + w_j)$. Indeed, for fixed $j \neq a$, every index $k \notin \{a, j\}$ is minimized at $\lambda_k = x_k$, so it remains to minimize

$$h_j(u,v) := w_a \,\mathrm{kl}(x_a, u) + w_j \,\mathrm{kl}(x_j, v)$$

over the convex constraint set $\{(u,v) \in (0,1)^2 : \ v \geq u\}$. If $w_a = w_j = 0$, then $h_j \equiv 0$ and the infimum is 0. If $w_a > 0$ and $w_j = 0$, the infimum is 0: take $u = x_a$ and any feasible $v \geq u$. If $w_a = 0$ and $w_j > 0$, the infimum is 0: take $v = x_j$ and any feasible $u \leq v$. It remains to consider $w_a, w_j > 0$. Since $x_a > x_j$, the unconstrained minimizer (x_a, x_j) is infeasible, and in this case h_j is strictly convex on $(0,1)^2$. Hence the constrained minimizer lies on the boundary $u = v$. Writing $u = v = m$, we minimize

$$m \longmapsto w_a \,\mathrm{kl}(x_a, m) + w_j \,\mathrm{kl}(x_j, m),$$

whose derivative is

$$\frac{w_a(m - x_a) + w_j(m - x_j)}{m(1-m)}.$$

This vanishes at the weighted average $m = m_{a,j}(x,w) \in (0,1)$, proving the formula above for g_j. Shrinking U if necessary, assume that $\overline{U} \subset (0,1)^K$ is compact. If $w_a + w_j > 0$, then $m_{a,j}(x,w)$ is a convex combination of x_a and x_j, so all arguments of the KL terms stay in a compact subset of $(0,1)$ and the formula above is continuous. If $(x^{(n)}, w^{(n)}) \to (x,w)$ with $w_a + w_j = 0$, then necessarily $w_a^{(n)} + w_j^{(n)} \to 0$; the corresponding KL terms remain uniformly bounded on $\overline{U}$, hence $g_j(x^{(n)}, w^{(n)}) \to 0 = g_j(x,w)$. Thus each g_j is continuous on $U \times \Delta_K$, and therefore so is the finite minimum $(x,w) \mapsto F(x,w)$.

By continuity of $(x,w) \mapsto F(x,w)$ at $(\mu, w^\star)$, there exist $\rho_0 > 0$ and $\eta \in (0, w_{\min}/2]$ such that

$$\|x - \mu\|_\infty \leq \rho_0 \text{ and } \|w - w^\star\|_\infty \leq \eta \quad \Longrightarrow \quad F(x,w) \geq (1-\varepsilon)\, F(\mu, w^\star) = \frac{1-\varepsilon}{T^\star(\mu)}.$$

Shrink ρ_0 further, if necessary, so that $B_\infty(\mu, \rho_0) \subset U \subset (0,1)^K$. Set

$$\rho := \min\{\rho_0, \Delta_{\min}/4\}.$$

We claim that whenever

$$\|\hat{\mu}(t) - \mu\|_\infty \leq \rho \quad \text{and} \quad \|N(t)/t - w^\star\|_\infty \leq \eta,$$

the conclusion (4) holds. On $\{\|\hat{\mu}(t) - \mu\|_\infty \leq \rho\}$, for every $j \neq a^\star$,

$$\hat{\mu}_{a^\star}(t) - \hat{\mu}_j(t) \geq (\mu_{a^\star} - \mu_j) - 2\rho \geq \Delta_{\min} - 2\rho \geq \Delta_{\min}/2 > 0,$$

so the empirical best arm is unique and equals $a^\star(\hat{\mu}(t)) = a^\star$. By Lemma 1, this implies $\Gamma_i(t) = 0$ for all $i \neq a^\star$, hence $J_t = a^\star$.

Moreover, since $a^\star(\hat{\mu}(t)) = a^\star$, we have $\mathrm{Alt}(\hat{\mu}(t)) = \mathrm{Alt}(\mu)$ and therefore

$$\Gamma_{a^\star}(t) = \inf_{\lambda \in \mathrm{Alt}(\mu)} \sum_{i=1}^{K} N_i(t)\, \mathrm{kl}(\hat{\mu}_i(t), \lambda_i) = t\, F\Big(\hat{\mu}(t), \frac{N(t)}{t}\Big).$$

Under the two conditions above this gives $\Gamma_{a^\star}(t) \geq \frac{1-\varepsilon}{T^\star(\mu)}\, t$. □

Proof of Theorem 3. Fix $\alpha \in (0,1)$ and write $w^\star := w^\star(\mu)$. Let $w_{\min} := \min_{1 \leq i \leq K} w_i^\star > 0$. Let $\rho > 0$ and $\eta \in (0, w_{\min}/2]$ be the constants from Lemma 4 applied with $\varepsilon = \alpha$. We use two random stabilization times to turn this deterministic implication into a bound on the expected stabilization time.

Step 1: reduce to integrability of two stabilization times. Define the allocation stabilization time

$$\tau_A := \inf\Big\{ t \geq K : \|N(s)/s - w^\star\|_\infty \leq \eta \text{ for all } s \geq t \Big\},$$

and define the time immediately after the last empirical failure while the allocation condition holds

$$\tau_B := 1 + \sup\Big\{ s \geq K : \|\hat{\mu}(s) - \mu\|_\infty > \rho,\ \|N(s)/s - w^\star\|_\infty \leq \eta \Big\},$$

with the convention $\sup \emptyset := K$. These times depend on the future trajectory and need not be stopping times; the argument below uses only their tail bounds and expectations. By the definition of τ_B, for every integer $s \geq \tau_B$, the allocation condition $\|N(s)/s - w^\star\|_\infty \leq \eta$ implies $\|\hat{\mu}(s) - \mu\|_\infty \leq \rho$. Together with the definition of τ_A, this implies that for any $t \geq \max\{K+1, \tau_A, \tau_B\}$, the conditions in (4) hold for every $s \geq t$, hence

$$t_\alpha \leq \max\{K+1, \tau_A, \tau_B\} \leq 1 + \max\{\tau_A, \tau_B\}.$$

Therefore,

$$\mathbb{E}_\mu[t_\alpha] \leq 1 + \mathbb{E}_\mu[\tau_A] + \mathbb{E}_\mu[\tau_B].$$

It remains to show that both expectations are finite.

Step 2: control $\mathbb{E}_\mu[\tau_B]$ and $\mathbb{E}_\mu[\tau_A]$. We first prove $\mathbb{E}_\mu[\tau_B] < \infty$ (Step 3), then $\mathbb{E}_\mu[\tau_A] < \infty$ (Step 4).

Step 3: empirical-mean term is exponentially small when the allocation is good. Whenever $\|N(s)/s - w^\star\|_\infty \leq \eta$, we have $N_i(s) \geq (w_{\min} - \eta)s \geq (w_{\min}/2)s$ for all i. Fix an arm i and set $n_s := \lceil (w_{\min}/2)s \rceil$. Then

$$\mathbb{P}_\mu\big(|\hat{\mu}_i(s) - \mu_i| > \rho,\ \|N(s)/s - w^\star\|_\infty \leq \eta\big) \leq \mathbb{P}_\mu\big(|\hat{\mu}_i(s) - \mu_i| > \rho,\ N_i(s) \geq n_s\big) \leq C_\rho e^{-2\rho^2 n_s} \leq c_1 e^{-c_2 s}$$

for some constants $c_1, c_2 > 0$. A union bound over i yields

$$\mathbb{P}_\mu\big(\|\hat{\mu}(s) - \mu\|_\infty > \rho,\ \|N(s)/s - w^\star\|_\infty \leq \eta\big) \leq c_1' e^{-c_2 s}.$$

As a consequence, for all large t,

$$\begin{aligned}
\mathbb{P}_\mu(\tau_B > t) &= \mathbb{P}_\mu\big(\exists s \geq t : \|\hat{\mu}(s) - \mu\|_\infty > \rho,\ \|N(s)/s - w^\star\|_\infty \leq \eta\big) \\
&\leq \sum_{s \geq t} \mathbb{P}_\mu\big(\|\hat{\mu}(s) - \mu\|_\infty > \rho,\ \|N(s)/s - w^\star\|_\infty \leq \eta\big) \\
&\leq c_1'' e^{-c_2 t}.
\end{aligned}$$

Thus $\mathbb{E}_\mu[\tau_B] < \infty$.

Step 4: allocation term. Write $w_t := w^\star(\theta_t)$ for $t \geq K+1$. For $t \geq K$, define the filtration

$$\mathcal{G}_t := \sigma(\mathcal{F}_t, \theta_{K+1}, \dots, \theta_t), \qquad \mathcal{G}_K = \mathcal{F}_K.$$

Then, for $t \geq K+1$,

$$\mathcal{G}_{t-1} \vee \sigma(\theta_t) = \sigma(\mathcal{F}_{t-1}, \theta_{K+1}, \dots, \theta_t), \qquad \mathbb{P}_\mu\big(I_t = i \mid \mathcal{G}_{t-1} \vee \sigma(\theta_t)\big) = w_{t,i}.$$

For each arm i, set

$$M_i(t) := N_i(t) - 1 - \sum_{u=K+1}^{t} w_{u,i}, \qquad t \geq K.$$

Then $M_i(K) = 0$, and $M_i(t)$ is $\mathcal{G}_t$-measurable. For $t \geq K+1$,

$$M_i(t) - M_i(t-1) = \mathbf{1}\{I_t = i\} - w_{t,i},$$

and the tower property gives

$$\begin{aligned}\mathbb{E}_\mu\big[M_i(t) - M_i(t-1) \mid \mathcal{G}_{t-1}\big] &= \mathbb{E}_\mu\big[\mathbb{E}_\mu\big[\mathbf{1}\{I_t = i\} - w_{t,i} \mid \mathcal{G}_{t-1} \vee \sigma(\theta_t)\big] \,\big|\, \mathcal{G}_{t-1}\big] \\ &= 0.\end{aligned}$$

Thus $(M_i(t))_{t \geq K}$ is a martingale with respect to $(\mathcal{G}_t)_{t \geq K}$, and its increments are bounded by 1. For any $x > 0$ and any $s \geq K+1$, Azuma–Hoeffding gives

$$\mathbb{P}\big(|M_i(s)| \geq x\big) \leq 2\exp\Big(-\frac{x^2}{2(s-K)}\Big) \leq 2\exp\Big(-\frac{x^2}{2s}\Big).$$

With $x = \sqrt{8s\log s}$ and a union bound over i,

$$\mathbb{P}\Big(\Big\|N(s) - \mathbf{1} - \sum_{u=K+1}^{s} w_u\Big\|_\infty \geq \sqrt{8s\log s}\Big) \leq 2K\, s^{-4}. \tag{5}$$

The same inequality is trivial at $s = K$ because $M_i(K) = 0$ for every arm i. By continuity of $\theta \mapsto w^\star(\theta)$ at μ, choose $r_0 \in (0, r_\mu]$ such that

$$\|\theta - \mu\|_\infty \leq r_0 \quad \Longrightarrow \quad \|w^\star(\theta) - w^\star\|_\infty \leq \eta/8. \tag{6}$$

For any $t \geq K+1$, if $\|\theta_t - \mu\|_\infty \leq r_0$ then (6) gives $\|w_t - w^\star\|_\infty \leq \eta/8$, while always $\|w_t - w^\star\|_\infty \leq 1$ since $w_t, w^\star \in \Delta_K$. Hence, for all $t \geq K+1$,

$$\|w_t - w^\star\|_\infty \leq \eta/8 + \mathbf{1}\{\|\theta_t - \mu\|_\infty > r_0\}.$$

Therefore, for any $s \geq K$,

$$\Big\|\frac{1}{s}\sum_{t=K+1}^{s} w_t - w^\star\Big\|_\infty \leq \frac{\eta}{8} + \frac{1}{s}\sum_{t=K+1}^{s} \mathbf{1}\{\|\theta_t - \mu\|_\infty > r_0\} + \frac{K}{s}.$$

Consequently,

$$\Big\|\frac{1}{s}\mathbf{1} + \frac{1}{s}\sum_{t=K+1}^{s} w_t - w^\star\Big\|_\infty \leq \frac{\eta}{8} + \frac{1}{s}\sum_{t=K+1}^{s} \mathbf{1}\{\|\theta_t - \mu\|_\infty > r_0\} + \frac{K+1}{s}.$$

Choose a deterministic s_0 so that $(K+1)/s \leq \eta/4$ and $\sqrt{8s\log s}/s \leq \eta/4$ for all $s \geq s_0$. Then for every

$$\begin{aligned} s \geq \max\Bigg\{ & s_0, \\ & \left\lceil \frac{8}{\eta} \sum_{t=K+1}^{\infty} \mathbf{1}\{\|\theta_t - \mu\|_\infty > r_0\} \right\rceil, \\ & 1 + \sup\left\{ u \geq K : \left\| N(u) - \mathbf{1} - \sum_{v=K+1}^{u} w_v \right\|_\infty \geq \sqrt{8u\log u} \right\} \Bigg\}, \end{aligned}$$

we have

$$\left\| \frac{N(s)}{s} - w^\star \right\|_\infty \le \eta.$$

Thus

$$\tau_A \le \max\Bigg\{ s_0, \left\lceil \frac{8}{\eta} \sum_{t=K+1}^{\infty} \mathbf{1}\{\|\theta_t - \mu\|_\infty > r_0\} \right\rceil, 1 + \sup\left\{ u \ge K : \left\| N(u) - \mathbf{1} - \sum_{v=K+1}^{u} w_v \right\|_\infty \ge \sqrt{8u \log u} \right\} \Bigg\},$$

with the convention $\sup \emptyset := K$. By (5) and a union bound,

$$\mathbb{P}_\mu\left(1 + \sup\left\{ u \ge K : \left\| N(u) - \mathbf{1} - \sum_{v=K+1}^{u} w_v \right\|_\infty \ge \sqrt{8u \log u} \right\} > t \right) \le \sum_{u \ge t} 2Ku^{-4} \le c\, t^{-3},$$

so the third term on the right-hand side has finite expectation. Taking expectations yields

$$\begin{aligned} \mathbb{E}_\mu[\tau_A] \le\ & s_0 + 1 \\ & + \mathbb{E}_\mu\left[1 + \sup\left\{ u \ge K : \left\| N(u) - \mathbf{1} - \sum_{v=K+1}^{u} w_v \right\|_\infty \ge \sqrt{8u \log u} \right\} \right] \\ & + \frac{8}{\eta}\, \mathbb{E}_\mu\left[\sum_{t=K+1}^{\infty} \mathbf{1}\{\|\theta_t - \mu\|_\infty > r_0\} \right]. \end{aligned}$$

By Tonelli's theorem,

$$\mathbb{E}_\mu\left[\sum_{t=K+1}^{\infty} \mathbf{1}\{\|\theta_t - \mu\|_\infty > r_0\} \right] = \sum_{t=K+1}^{\infty} \mathbb{P}_\mu\big(\|\theta_t - \mu\|_\infty > r_0\big).$$

We use the following tail bound of the posterior draw.

Lemma 5 (Polynomial tail bound for posterior draws)**.** *For every $r_0 \in (0, r_\mu]$, there exist $c_3(r_0) > 0$, $p(r_0) > 1$, and $t_0(r_0) \ge 3$ such that*

$$\mathbb{P}_\mu(\|\theta_t - \mu\|_\infty > r_0) \le c_3(r_0)\, t^{-p(r_0)} \qquad \textit{for all } t \ge t_0(r_0).$$

The full proof of Lemma 5 is given in Appendix G.

Applying Lemma 5, there exist $c_3(r_0) > 0$, $p(r_0) > 1$, and $t_0(r_0) \ge 3$ such that

$$\mathbb{P}_\mu(\|\theta_t - \mu\|_\infty > r_0) \le c_3(r_0)\, t^{-p(r_0)} \qquad \text{for all } t \ge t_0(r_0).$$

Since $p(r_0) > 1$, the above series converges, and therefore $\mathbb{E}_\mu[\tau_A] < \infty$.

Conclusion. Step 3 gives $\mathbb{E}_\mu[\tau_B] < \infty$ and Step 4 gives $\mathbb{E}_\mu[\tau_A] < \infty$. Together with Step 1, this yields $\mathbb{E}_\mu[t_\alpha] < \infty$. □

F Proof of summable logarithmic exploration

Recall the fixed radius $r_\mu > 0$ from (3) and define, for $t \ge K$, the good empirical event

$$\mathcal{E}_t := \bigcap_{i=1}^{K} \Big\{ |\hat{\mu}_i(t) - \mu_i| \le r_\mu/2 \Big\}.$$

The goal of this section is to prove that every arm is sampled at least logarithmically often, with a summable tail bound. The key probabilistic inputs are that, once all pull counts are already large, failures of $\mathcal{E}_t$ are exponentially rare, and that the probability of a long non-interruption block is polynomially small. The long-gap reduction below then converts these ingredients into Proposition 1.

The next lemma is the stopping-time version of the usual Hoeffding control for empirical means. Later we will apply it at random times such as the start of a long non-interruption block.

Lemma 6 (Hoeffding control of $\mathcal{E}_\tau^c$ at large pull counts)**.** *Fix an arm $i \in \{1, \dots, K\}$ and $u > 0$. For any (possibly random) time τ, and any integer $n \geq 1$,*

$$\mathbb{P}_\mu\big(|\hat{\mu}_i(\tau) - \mu_i| > u,\ N_i(\tau) \geq n\big) \leq C_u \exp\big(-2u^2 n\big), \qquad C_u := \frac{2}{1 - e^{-2u^2}}.$$

In particular, taking $u = r_\mu/2$,

$$\mathbb{P}_\mu\big(\mathcal{E}_\tau^c,\ \min_{1 \leq i \leq K} N_i(\tau) \geq n\big) \leq K C_{r_\mu/2} \exp\Big(-\frac{n r_\mu^2}{2}\Big), \qquad C_{r_\mu/2} := \frac{2}{1 - e^{-r_\mu^2/2}}.$$

Proof. Fix $n \geq 1$. Let $(Y_{i,\ell})_{\ell \geq 1}$ denote the rewards obtained from the successive pulls of arm i. Under $\mathbb{P}_\mu$, these are i.i.d. $\mathrm{Bern}(\mu_i)$. For any integer $m \geq 1$, on the event $\{N_i(\tau) = m\}$ we have

$$\hat{\mu}_i(\tau) = \frac{1}{m} \sum_{\ell=1}^{m} Y_{i,\ell}.$$

This is why the random time causes no extra difficulty: once we condition on the event $\{N_i(\tau) = m\}$, the empirical mean is exactly the average of the first m rewards from arm i. Hence, for every $m \geq n$,

$$\big\{|\hat{\mu}_i(\tau) - \mu_i| > u,\ N_i(\tau) = m\big\} \subseteq \Big\{\Big|\frac{1}{m} \sum_{\ell=1}^{m} Y_{i,\ell} - \mu_i\Big| > u\Big\}.$$

By Hoeffding's inequality, $\mathbb{P}_\mu\big(|\hat{\mu}_i(\tau) - \mu_i| > u,\ N_i(\tau) = m\big) \leq 2e^{-2u^2 m}$. Summing over $m \geq n$ and using the geometric series gives

$$\mathbb{P}_\mu\big(|\hat{\mu}_i(\tau) - \mu_i| > u,\ N_i(\tau) \geq n\big) \leq \sum_{m \geq n} 2e^{-2u^2 m} = \frac{2e^{-2u^2 n}}{1 - e^{-2u^2}} = C_u e^{-2u^2 n}.$$

The last equality is just the geometric-series formula. The second claim follows by a union bound over i with $u = r_\mu/2$. □

We now prove Proposition 1; the proof is given in Appendix F.1.

F.1 Long-gap proof of Proposition 1

Proof of Proposition 1. We follow the long-gap strategy of Kaufmann et al. (2012), adapted to posterior tracking. The picture is simple: if arm i has been pulled fewer than $A \log t$ times by time t, then one of the gaps between consecutive pulls of i must be long. That long gap contains only finitely many interruptions, so one of the resulting non-interruption blocks is still long. The deterministic reduction from a long i-gap to such a block is recorded below in Lemma 9; the only probabilistic input is Lemma 10.

Fix an arm i and a horizon t. Let $\tau_{i,0} := 0$, and for $j \geq 1$ let $\tau_{i,j}$ be the time of the j-th play of arm i, with the convention $\tau_{i,j} := t + 1$ if arm i is played fewer than j times by time t. Set

$$L := \big\lfloor t/(A \log t + 1) \big\rfloor - 1, \qquad E_{i,j} := \big\{\tau_{i,j+1} - \tau_{i,j} - 1 \geq L\big\}.$$

If $N_i(t) < A \log t$ and $n := N_i(t)$, then $\tau_{i,n+1} = t + 1$ and

$$\sum_{j=0}^{n} \big(\tau_{i,j+1} - \tau_{i,j} - 1\big) = t - n.$$

If every one of these $n+1$ gaps had length at most $L-1$, then

$$t - n \le (n+1)(L-1) < (A\log t + 1)L \le t - n,$$

a contradiction. Hence some $E_{i,j}$ occurs, and therefore

$$\mathbb{P}_\mu\big(N_i(t) < A\log t\big) \le \sum_{0 \le j \le \lceil A\log t\rceil - 1} \mathbb{P}_\mu(E_{i,j}). \tag{7}$$

This is simply a counting argument: if every gap were short, then the total length of all gaps would be too small to fill the interval $\{1,\dots,t\}$.

Step 1: probabilistic inputs. This step fixes the thresholds and comparison quantities that will later control the escape probability from a long block. Recall from Appendix C that

$$\Delta_{\min} := \mu_{a^\star(\mu)} - \max_{j \ne a^\star(\mu)} \mu_j > 0.$$

Recall that the fixed radius r_μ from (3) satisfies $r_\mu \le \Delta_{\min}/8$. Set

$$\mu^-_{a^\star(\mu)} := \mu_{a^\star(\mu)} - \frac{\Delta_{\min}}{4} \in (0,1), \qquad R_\mu := \inf_{x \in [0, \mu^-_{a^\star(\mu)})} \frac{\mathrm{kl}(x, \mu_{a^\star(\mu)})}{\mathrm{kl}(x, \mu^-_{a^\star(\mu)})}.$$

The quantity R_μ is strictly larger than 1.

Lemma 7. *One has $R_\mu > 1$.*

The proof of Lemma 7 is given in Appendix F.1.3. We write

$$\mathrm{kl}_-(x,u) := \begin{cases} \mathrm{kl}(x,u), & x \le u, \\ 0, & x > u. \end{cases}$$

Step 2: saturation and posterior concentration on saturated arms. This step supplies the stopping-time posterior tail bound for arms that have already accumulated at least $A\log t$ samples.

Definition 1 (Saturated arms)**.** *For the fixed parameter A from the standing assumption, arm k is said to be saturated at time s if $N_k(s) \ge A\log t$. We write*

$$\mathcal{S}_s := \big\{k \in \{1,\dots,K\} : N_k(s) \ge A\log t\big\}.$$

We record the following posterior deviation bound on saturated arms (proof deferred to Subsubsection F.1.4). It is the stopping-time analogue of the fixed-sample posterior tail estimate proved earlier.

Lemma 8 (Posterior deviation on saturated arms)**.** *Fix the horizon t from Proposition 1. There exists a constant $c_0 = c_0(\mu) > 0$ such that for every $A \ge 1$ there are constants $C_1 = C_1(\mu, A) > 0$ and $t_1 = t_1(\mu, A) \ge 3$ with the following property: for every arm $k \in \{1,\dots,K\}$, every stopping time τ with $K \le \tau < \infty$ almost surely, and every $t \ge t_1$,*

$$\mathbb{P}_\mu\big(|\theta_{k,\tau+1} - \mu_k| > r_\mu,\ N_k(\tau) \ge A\log t\big) \le C_1\, t^{-Ac_0}.$$

Standing assumption. Fix $A \ge 1$ and assume throughout that t is large enough so that all bounds used below hold; the required lower threshold may increase from line to line. In particular, we take t large enough so that the explicit lower bounds on L' derived below hold when L' and M are defined.

Step 3: bounding a long gap. This step is deterministic: once a long gap is fixed, it is decomposed into blocks separated by interruptions, and one of those blocks must still be long. Fix an integer j satisfying $0 \le j \le \lceil A\log t\rceil - 1$ and define the gap interval

$$I := \{\tau_{i,j} + 1, \dots, \tau_{i,j+1} - 1\}.$$

On I, arm i is never sampled.

Definition 2 (Interruptions on the gap interval I)**.** *A time $s \in I$ is called an* interruption *if the algorithm pulls some arm $k \ne i$ with $N_k(s-1) < A\log t$ at time s.*

Moreover, on $E_{i,j}$ we have $|I| = (\tau_{i,j+1} - 1) - \tau_{i,j} \geq L$.

Interruptions and their total number. Each interruption increases the count of some arm $k \neq i$ by one, and once an arm becomes saturated it stays saturated. Hence the total number of interruptions on I is at most

$$M := (K-1)\lceil A \log t \rceil.$$

Set

$$L' := \left\lfloor \frac{L-M}{M+1} \right\rfloor.$$

The next lemma is purely deterministic. It says that if there are at most M interruptions inside a gap of length at least L, then at least one non-interruption block must have length at least L'.

Lemma 9 (Deterministic reduction from a long gap to a long block)**.** *Set $T_0 := \tau_{i,j}$ and, for $q \geq 0$,*

$$T_{q+1} := \Big(\inf\{s \in I : \ s > T_q \text{ and } s \text{ is an interruption}\}\Big) \wedge (t+1) \wedge \tau_{i,j+1}.$$

Define

$$B_q := \{T_q + 1, \dots, T_{q+1} - 1\}, \qquad \widetilde{B}_q := \{T_q + 1, \dots, T_q + L'\}, \qquad \widetilde{E}_{i,j,q} := \{|B_q| \geq L'\}.$$

Then each T_q is a bounded stopping time. Moreover, for every $q \in \{0, \dots, M\}$,

$$\widetilde{E}_{i,j,q} \subseteq \{T_q \geq K\}, \qquad E_{i,j} \subseteq \bigcup_{q=0}^{M} \widetilde{E}_{i,j,q}.$$

Proof. The first interruption time after a bounded stopping time is again a bounded stopping time, so the claim for (T_q) follows by induction. If $T_q < K$, then either $T_q + 1 = \tau_{i,j+1}$, or $T_q + 1 \in I \cap \{1, \dots, K\}$ and the arm pulled at time $T_q + 1$ is still unsaturated, hence $T_q + 1$ itself is an interruption. In both cases $T_{q+1} \in \{T_q + 1, \tau_{i,j+1}\}$, so $B_q = \varnothing$ and therefore $\widetilde{E}_{i,j,q} \subseteq \{T_q \geq K\}$.

On $E_{i,j}$, the interval I has length at least L and contains at most M interruptions. Hence it is split into at most $M+1$ non-interruption blocks $B_0, \dots, B_M$ with total length at least $L - M$. If all of them had length at most $L' - 1$, then

$$L - M \leq \sum_{q=0}^{M} |B_q| \leq (M+1)(L'-1) < L - M,$$

a contradiction. Thus some block has length at least L', which gives the covering of $E_{i,j}$. □

The next lemma is the only probabilistic estimate in the long-gap reduction: it shows that a long non-interruption block is itself polynomially rare.

Lemma 10 (Long non-interruption block bound)**.** *Let c_0 be the constant from Lemma 8. For any*

$$A \geq \max\Big\{1, \ \frac{2}{c_0}\Big\},$$

there exist constants $\gamma = \gamma(\mu, A) > 1$, $C = C(\mu, A) > 0$, and $t_0 = t_0(\mu, A) \geq 3$ such that for all $t \geq t_0$, all arms i, all $j \in \{0, \dots, \lceil A \log t \rceil - 1\}$, and all $q \in \{0, \dots, M\}$,

$$\mathbb{P}_\mu(\widetilde{E}_{i,j,q}) \leq C\, t^{-\gamma}.$$

The proof is given in Subsubsection F.1.2 below. Fix any A satisfying the assumption of Lemma 10, and let $\gamma > 1$, $C > 0$, and $t_0 \geq 3$ be the corresponding constants from Lemma 10. Choose any number

$$1 < c < \gamma.$$

For a fixed arm i, Lemma 9 and (7) give, for all large enough t,

$$\begin{aligned}\mathbb{P}_\mu\big(N_i(t) < A\log t\big) &\le \sum_{0\le j\le \lceil A\log t\rceil - 1}\ \sum_{q=0}^{M} \mathbb{P}_\mu(\widetilde{E}_{i,j,q}) \\ &\le C(A\log t + 1)(M+1)\, t^{-\gamma} \\ &\le C_1\,(A\log t)^2\, t^{-\gamma}\end{aligned}$$

for a constant $C_1 = C_1(\mu, A) > 0$. Since $c < \gamma$, after enlarging thresholds we may absorb the factor $(A\log t)^2$ into $t^{\gamma-c}$ and obtain

$$\mathbb{P}_\mu\big(N_i(t) < A\log t\big) \le C_2\, t^{-c}$$

for a constant $C_2 = C_2(\mu, A) > 0$. Union bounding over $i \in \{1, \dots, K\}$ and absorbing the factor K into the constant completes Proposition 1 (after enlarging thresholds). □

F.1.1 Auxiliary ingredients for Lemma 10

We record the auxiliary statements used in the proof of Lemma 10. Set

$$\delta_t := (\log t)^{-4}.$$

The four ingredients play different roles: the first freezes the unsaturated posteriors while keeping their current draws conditionally independent, the second turns separation of the current draw into a lower bound on oracle mass assigned to the unsaturated set, the third lower-bounds a fixed threshold tail of a Beta posterior, and the fourth shows that the only dangerous small-Φ_q event is itself polynomially rare.

The first lemma formalizes the key mechanism behind the escape argument: the posterior parameters of the unsaturated arms stay fixed, while the current unsaturated draws remain conditionally independent.

Lemma 11 (Unsaturated-arm anti-concentration under fixed posterior parameters)**.** *Fix i, j, and q in the setting of Lemma 10. Set*

$$\mathcal{U} := \{1, \dots, K\} \setminus \mathcal{S}_{T_q}.$$

For each $r \in \{1, \dots, L'\}$, write

$$s := T_q + r, \qquad \mathcal{H}_{q,r} := \sigma\Big(\mathcal{F}_{T_q+r-1}, \theta_{K+1}, \dots, \theta_{T_q+r-1}, \big(\theta_{k,T_q+r}\big)_{k\in\mathcal{S}_{T_q}}\Big).$$

Then, on Ξ_{r-1}, the posterior draws of the unsaturated arms at time s are conditionally independent given $\mathcal{H}_{q,r}$, and each arm u satisfies

$$\theta_{u,s} \mid \mathcal{H}_{q,r} \sim \mathrm{Beta}\big(\alpha_u + S_u(T_q), \beta_u + N_u(T_q) - S_u(T_q)\big) \qquad \text{a.s. on } \Xi_{r-1} \cap \{u \in \mathcal{U}\}.$$

If $a \in \{1, \dots, K\}$, then after conditioning on $\theta_{a,s}$ the remaining unsaturated arms are still conditionally independent: for each arm $u \neq a$,

$$\theta_{u,s} \mid \mathcal{H}_{q,r}, \theta_{a,s} \sim \mathrm{Beta}\big(\alpha_u + S_u(T_q), \beta_u + N_u(T_q) - S_u(T_q)\big) \qquad \text{a.s. on } \Xi_{r-1} \cap \{a, u \in \mathcal{U}\}.$$

Consequently, there exists a constant $C_0 > 0$, depending only on the prior, such that for all large enough t, all $r \in \{1, \dots, L'\}$, and all distinct arms u, a,

$$\sup_{x\in(0,1)} \mathbb{P}_\mu\big(|\theta_{u,s} - x| \le \delta_t \mid \mathcal{H}_{q,r}\big) \le \frac{C_0(1+A)}{(\log t)^3} \qquad \text{a.s. on } \Xi_{r-1} \cap \{u \in \mathcal{U}\}. \tag{8}$$

The same bound holds with conditioning on $\mathcal{H}_{q,r}, \theta_{a,s}$ and on the smaller event $\Xi_{r-1} \cap \{a, u \in \mathcal{U}\}$.

The next lemma is the deterministic oracle-allocation input: once the best and second-best arm values are interior and separated, the oracle must place noticeable mass on any nonempty candidate set U.

Lemma 12 (Oracle escape mass on a separated interior draw)**.** *Let $x \in (0,1)^K$ and let $U \subseteq \{1, \dots, K\}$ be nonempty. Let a be a best arm of x and let b be a second-best arm of x. Assume that $x_a, x_b \in [1/t, 1 - 1/t]$ and that $x_a - x_b \ge \delta_t$. Then*

$$\sum_{u\in U} w^\star_u(x) \ge \frac{1}{2K(\log t)^9}. \tag{9}$$

The next bound prevents the best-arm posterior from assigning too little mass above the fixed threshold τ.

Lemma 13 (Beta posterior lower tail mass above a fixed threshold)**.** *Fix $\tau \in (0,1)$ and integers $\alpha, \beta \geq 1$. There exists a constant $c_\tau = c_\tau(\tau, \alpha, \beta) > 0$ such that for all $n \geq 0$ and all $s \in \{0, \dots, n\}$, if*

$$\theta \sim \mathrm{Beta}(\alpha + s, \beta + n - s),$$

then

$$\mathbb{P}(\theta \geq \tau) \geq c_\tau \, (n+1)^{-1/2} \exp\Big(-n \, \mathrm{kl}_-(s/n, \tau)\Big),$$

where the exponent is interpreted as 0 when $n = 0$.

The final ingredient bounds the conditional posterior probability Φ_q from below on the event $\mathcal{M}_q$.

Lemma 14 (Lower bound for Φ_q on $\mathcal{M}_q$)**.** *Fix i, j, and q in the setting of Lemma 10. When $a^\star(\mu) \notin \mathcal{S}_{T_q}$, the escape estimate in Lemma 10 is governed by the conditional posterior mass of arm $a^\star(\mu)$ above the threshold $\mu^-_{a^\star(\mu)}$. Define*

$$\Phi_q := \mathbb{P}_\mu\Big(\theta_{a^\star(\mu), T_q + 1} \geq \mu^-_{a^\star(\mu)} \;\Big|\; \mathcal{F}_{T_q}\Big).$$

Let

$$\mathcal{M}_q := \Big\{K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \notin \mathcal{S}_{T_q},\ N_{a^\star(\mu)}(T_q)\, \mathrm{kl}_-\Big(\frac{S_{a^\star(\mu)}(T_q)}{N_{a^\star(\mu)}(T_q)}, \mu^-_{a^\star(\mu)}\Big) \leq \frac{1 + R_\mu}{2R_\mu} \log t\Big\}.$$

Then there exists a constant $c = c(\mu, A) > 0$ such that, almost surely on $\mathcal{M}_q$,

$$\Phi_q \geq c \, (A \log t + 1)^{-1/2} t^{-(1+R_\mu)/(2R_\mu)}.$$

The proof of Lemma 11 is given in Subsubsection F.1.5, the proof of Lemma 12 is given in Subsubsection F.1.6, the proof of Lemma 14 is given in Subsubsection F.1.8, and the proof of Lemma 13 is given in Subsubsection F.1.7.

F.1.2 Proof of Lemma 10

Proof of Lemma 10. Let c_0 be the constant from Lemma 8. Fix any

$$A \geq \max\Big\{1,\ \frac{2}{c_0}\Big\},$$

let $C_1 = C_1(\mu, A) > 0$ be the constant from Lemma 8, and then fix i, j, q. The lower bound on A ensures $Ac_0 \geq 2$, so the saturated-tail error term will already be summable in t. Set

$$\widetilde{B}_q := \{T_q + 1, \dots, T_q + L'\}.$$

If $T_q = \tau_{i,j+1}$, then by definition of B_q we have $B_q = \varnothing$, so $\widetilde{E}_{i,j,q}$ cannot occur. Together with the earlier inclusion $\widetilde{E}_{i,j,q} \subseteq \{T_q \geq K\}$, this implies

$$\widetilde{E}_{i,j,q} \subseteq \{K \leq T_q < \tau_{i,j+1}\}.$$

On $\widetilde{E}_{i,j,q}$ we have $\widetilde{B}_q \subseteq B_q \subseteq I$, and in particular arm i is never pulled on $\widetilde{B}_q$ (by the definition of the i-gap interval I).

Let $\mathcal{U} := \{1, \dots, K\} \setminus \mathcal{S}_{T_q}$ be the unsaturated set at time T_q. Since $T_0 = \tau_{i,j}$ and the sequence (T_q) is nondecreasing, on $\{K \leq T_q < \tau_{i,j+1}\}$ we also have $\tau_{i,j} \leq T_q < \tau_{i,j+1}$. By the definition of $\tau_{i,j+1}$ as the time of the $(j+1)$-st play of arm i, this means $N_i(T_q) = j$. Since in Step 3 we fixed $j \in \{0, \dots, \lceil A \log t \rceil - 1\}$, it follows that $N_i(T_q) = j < A \log t$, so

$$i \in \mathcal{U}, \qquad i \notin \mathcal{S}_{T_q}, \qquad \mathcal{U} \neq \varnothing.$$

This nonemptiness matters because the escape mechanism below acts through $\mathcal{U}$: later we will use anti-concentration on the unsaturated arms together with a lower bound on the oracle mass assigned to $\mathcal{U}$.

At the level of orders, L has size $t/(A\log t)$ while M has size $KA\log t$, so one expects L' to have size $t/(KA^2(\log t)^2)$. Concretely, for all large enough t,

$$L' = \Big\lfloor \frac{L-M}{M+1} \Big\rfloor \geq \frac{1}{16KA^2}\,\frac{t}{(\log t)^2}.$$

Fix t large enough so that $\delta_t \leq \Delta_{\min}/4$ and $L' \geq 4$. At finitely many later steps we may enlarge this lower threshold on t further, without changing notation.

For $\ell = 0, 1, \ldots, L'$, let Ξ_ℓ denote the event that, starting from time T_q, the algorithm spends the next ℓ rounds sampling only arms that were already saturated at time T_q; equivalently, no arm from the unsaturated set $\mathcal{U} = \{1,\ldots,K\} \setminus \mathcal{S}_{T_q}$ is sampled during $\{T_q+1,\ldots,T_q+\ell\}$. We define

$$\Xi_\ell := \{K \leq T_q < \tau_{i,j+1}\} \cap \bigcap_{r=1}^{\ell} \{I_{T_q+r} \in \mathcal{S}_{T_q}\}.$$

In particular,

$$\Xi_0 = \{K \leq T_q < \tau_{i,j+1}\}, \qquad \Xi_r = \Xi_{r-1} \cap \{I_{T_q+r} \in \mathcal{S}_{T_q}\} \quad \text{for } r = 1,\ldots,L'. \tag{10}$$

Thus $\Xi_{L'}$ is the event that during $\widetilde{B}_q = \{T_q+1,\ldots,T_q+L'\}$ the algorithm only pulls arms in $\mathcal{S}_{T_q}$.

The proof is an escape-from-saturation argument. By Lemma 9, a long i-gap yields a long interruption-free block. On such a block, staying inside $\mathcal{S}_{T_q}$ for L' consecutive rounds means repeatedly failing to sample from $\mathcal{U}$. We therefore prove that, except when too many saturated draws are bad, there is an explicit lower bound on the conditional probability of sampling from $\mathcal{U}$ at each good round, and iterating this one-step estimate gives exponential decay. The two cases $a^\star(\mu) \in \mathcal{S}_{T_q}$ and $a^\star(\mu) \in \mathcal{U}$ differ only in the lower bound used for that one-step probability.

1. Reduction to $\Xi_{L'}$. We now work on the event $\widetilde{E}_{i,j,q}$. The first task is to prove

$$\widetilde{E}_{i,j,q} \subseteq \Xi_{L'},$$

that is, to show that on $\widetilde{E}_{i,j,q}$ every round in $\widetilde{B}_q$ pulls an arm in $\mathcal{S}_{T_q}$. Suppose therefore that $\widetilde{E}_{i,j,q}$ occurs and that some arm $u \in \mathcal{U}$ is pulled on $\widetilde{B}_q$; let s be the first such round. Since $i \in \mathcal{U}$ and arm i is never pulled on $\widetilde{B}_q \subseteq I$, we have $u \neq i$. By the minimality of s, no arm in $\mathcal{U}$ has been pulled before time s, so u is still unsaturated at time $s-1$. Thus s would be an interruption, contradicting $\widetilde{B}_q \subseteq B_q$. This proves $\widetilde{E}_{i,j,q} \subseteq \Xi_{L'}$.

Two consequences will be used repeatedly below. First, since $\Xi_{L'} \subseteq \{K \leq T_q < \tau_{i,j+1}\}$ by definition,

$$\mathbb{P}_\mu(\widetilde{E}_{i,j,q}) \leq \mathbb{P}_\mu(\Xi_{L'}).$$

Second, on $\widetilde{E}_{i,j,q}$ no arm in $\mathcal{U}$ is pulled on $\widetilde{B}_q$, so the pull counts of arms in $\mathcal{U}$ do not change on $\widetilde{B}_q$; meanwhile every arm already in $\mathcal{S}_{T_q}$ remains saturated throughout the block by monotonicity.

For each $s \in \widetilde{B}_q$, define the saturated-good event

$$\mathrm{Good}_s := \bigcap_{k \in \mathcal{S}_{T_q}} \big\{|\theta_{k,s} - \mu_k| \leq r_\mu\big\}, \tag{11}$$

and let

$$\mathrm{Bad} := \sum_{s \in \widetilde{B}_q} \mathbf{1}\{\mathrm{Good}_s^c\}.$$

Thus Good_s is the event that $|\theta_{k,s} - \mu_k| \leq r_\mu$ for every $k \in \mathcal{S}_{T_q}$, while Bad counts the rounds on which this event fails. The event Good_s only constrains the arms in $\mathcal{S}_{T_q}$: those arms will be controlled via posterior concentration, while the arms in $\mathcal{U}$ will be handled later through anti-concentration and a lower bound on the oracle mass assigned to unsaturated arms. Using the inclusion $\widetilde{E}_{i,j,q} \subseteq \Xi_{L'}$, we

now split according to whether more than half of the rounds in $\widetilde{B}_q$ are bad and, on the complementary event, according to whether the true best arm lies in $\mathcal{S}_{T_q}$ or in $\mathcal{U}$. This gives

$$\begin{aligned}\mathbb{P}_\mu(\widetilde{E}_{i,j,q}) \le{}& \underbrace{\mathbb{P}_\mu\big(\mathrm{Bad} > L'/2,\ K \le T_q < \tau_{i,j+1}\big)}_{\text{(a)}} \\ &+ \underbrace{\mathbb{P}_\mu\big(\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\}\big)}_{\text{(b)}} \\ &+ \underbrace{\mathbb{P}_\mu\big(\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\} \cap \{a^\star(\mu) \in \mathcal{U}\}\big)}_{\text{(c)}}.\end{aligned}$$

For term (a) we dropped $\Xi_{L'}$ and kept only the coarser event $\{K \le T_q < \tau_{i,j+1}\}$, because the saturated-tail estimate in Step 2 does not use the event $\Xi_{L'}$. Terms (b) and (c) retain $\Xi_{L'}$, since the later block argument does use it.

2. Bounding term (a). Fix $\ell \in \{1, \dots, L'\}$. Set $\tau := (T_q + \ell - 1) \vee K$. Since T_q is a bounded stopping time by Lemma 9, so is τ, and $\tau \ge K$ almost surely. On $\{K \le T_q < \tau_{i,j+1}\}$, we have $\tau = T_q + \ell - 1$. On $\{K \le T_q < \tau_{i,j+1}\}$, since $i \notin \mathcal{S}_{T_q}$,

$$\begin{aligned}\{K \le T_q < \tau_{i,j+1}\} \cap \mathrm{Good}^c_{T_q+\ell} &= \{K \le T_q < \tau_{i,j+1}\} \cap \bigcup_{k \in \mathcal{S}_{T_q}} \big\{|\theta_{k,T_q+\ell} - \mu_k| > r_\mu\big\} \\ &= \{K \le T_q < \tau_{i,j+1}\} \cap \bigcup_{k \ne i} \Big(\{k \in \mathcal{S}_{T_q}\} \cap \big\{|\theta_{k,T_q+\ell} - \mu_k| > r_\mu\big\}\Big).\end{aligned}$$

On $\{K \le T_q < \tau_{i,j+1}\} \cap \{k \in \mathcal{S}_{T_q}\}$, we have $N_k(T_q) \ge A \log t$ by definition of $\mathcal{S}_{T_q}$. Since $N_k(\cdot)$ is nondecreasing and $\tau = T_q + \ell - 1$ on this event, it follows that $N_k(\tau) \ge A \log t$. Hence

$$\begin{aligned}&\{K \le T_q < \tau_{i,j+1}\} \cap \{k \in \mathcal{S}_{T_q}\} \cap \big\{|\theta_{k,T_q+\ell} - \mu_k| > r_\mu\big\} \\ &\quad\subseteq \big\{|\theta_{k,\tau+1} - \mu_k| > r_\mu,\ N_k(\tau) \ge A \log t\big\}.\end{aligned}$$

Therefore, by a union bound and Lemma 8 applied at τ,

$$\begin{aligned}\mathbb{P}_\mu\big(\mathrm{Good}^c_{T_q+\ell},\ K \le T_q < \tau_{i,j+1}\big) &\le \sum_{k \ne i} \mathbb{P}_\mu\big(|\theta_{k,\tau+1} - \mu_k| > r_\mu,\ N_k(\tau) \ge A \log t\big) \\ &\le (K-1) C_1\, t^{-Ac_0}.\end{aligned}$$

Since ℓ was arbitrary, the same bound holds for every $\ell \in \{1, \dots, L'\}$. Hence

$$\mathbb{E}_\mu[\mathrm{Bad}\, \mathbf{1}\{K \le T_q < \tau_{i,j+1}\}] \le L'(K-1) C_1 t^{-Ac_0},$$

and therefore

$$\begin{aligned}(a) &= \mathbb{P}_\mu\big(\mathrm{Bad} > L'/2,\ K \le T_q < \tau_{i,j+1}\big) \\ &\le \frac{2}{L'}\, \mathbb{E}_\mu[\mathrm{Bad}\, \mathbf{1}\{K \le T_q < \tau_{i,j+1}\}] \\ &\le 2(K-1) C_1\, t^{-Ac_0}.\end{aligned} \tag{12}$$

3. Bounding term (b). We now turn to the complementary event $\{\mathrm{Bad} \le L'/2\}$. On $\{\mathrm{Bad} \le L'/2\} \cap \{K \le T_q < \tau_{i,j+1}\}$, at least $\lceil L'/2 \rceil$ of the rounds in $\widetilde{B}_q$ are good. For term (b), we use those good rounds in three stages: we first prove a one-step escape bound on $\Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell}$, then iterate that estimate along the block by the chain rule, and finally use that at least $\lceil L'/2 \rceil$ rounds are good to obtain exponential decay. In other words, to bound

$$\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\},$$

we first study, for each ℓ, the local event

$$\begin{aligned}&\Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\} \\ &\quad= \{K \le T_q < \tau_{i,j+1}\} \cap \bigcap_{m=1}^{\ell-1} \{I_{T_q+m} \in \mathcal{S}_{T_q}\} \cap \bigcap_{k \in \mathcal{S}_{T_q}} \{|\theta_{k,T_q+\ell} - \mu_k| \le r_\mu\} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\},\end{aligned}$$

and then combine those one-step bounds across the block.

(b)-1. One-step escape bound.

Fix $\ell \in \{1, \dots, L'\}$ and let $\mathcal{H}_{q,\ell}$ be the sigma-field from Lemma 11. Then $\Xi_{\ell-1}$ and $\mathrm{Good}_{T_q+\ell}$ are $\mathcal{H}_{q,\ell}$-measurable. On $\Xi_{\ell-1}$, no arm in $\mathcal{U}$ is sampled during $\{T_q+1, \dots, T_q+\ell-1\}$, so the unsaturated posterior parameters remain equal to their values at time T_q. Hence, conditional on $\mathcal{H}_{q,\ell}$, the current posterior draws of the unsaturated arms at round $T_q+\ell$ are independent and satisfy the one-dimensional and pairwise anti-concentration bounds from Lemma 11. In particular, $\mathcal{H}_{q,\ell}$ reveals the past and the current draws of the saturated arms, but not the current draws of the unsaturated arms.

To obtain the one-step escape bound for term (b), we now work under the conjunction

$$\Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\}.$$

On this event, every $k \in \mathcal{S}_{T_q} \setminus \{a^\star(\mu)\}$ satisfies

$$\theta_{a^\star(\mu),T_q+\ell} - \theta_{k,T_q+\ell} \geq (\mu_{a^\star(\mu)} - r_\mu) - (\mu_k + r_\mu) \geq \Delta_{\min} - 2r_\mu \geq \frac{3\Delta_{\min}}{4} > \delta_t. \tag{13}$$

Moreover, on $\mathrm{Good}_{T_q+\ell}$ the posterior draw of every saturated arm lies in $[1/t, 1-1/t]$ for all large enough t. We now define an auxiliary event $G_{T_q+\ell}$ on which the posterior draw of every unsaturated arm lies in $[1/t, 1-1/t]$, the posterior draws of distinct unsaturated arms are more than δ_t apart, and every unsaturated arm draw is more than δ_t away from $\theta_{a^\star(\mu),T_q+\ell}$:

$$\begin{aligned} G_{T_q+\ell} := &\Big\{\theta_{u,T_q+\ell} \in [1/t, 1-1/t] \text{ for all } u \in \mathcal{U}\Big\} \\ &\cap \bigcap_{\substack{u,v\in\mathcal{U}\\ u\neq v}} \{|\theta_{u,T_q+\ell} - \theta_{v,T_q+\ell}| > \delta_t\} \cap \bigcap_{u\in\mathcal{U}} \{|\theta_{u,T_q+\ell} - \theta_{a^\star(\mu),T_q+\ell}| > \delta_t\}. \end{aligned}$$

By (8) and by the pairwise anti-concentration bound obtained from Lemma 11 by conditional independence, and since $1/t \leq \delta_t$, we have

$$\{\theta_{u,T_q+\ell} \notin [1/t, 1-1/t]\} \subseteq \{|\theta_{u,T_q+\ell} - 1/t| \leq \delta_t\} \cup \{|\theta_{u,T_q+\ell} - (1-1/t)| \leq \delta_t\}$$

for every $u \in \mathcal{U}$. Therefore a union bound gives

$$\begin{aligned} \mathbb{P}_\mu(G^c_{T_q+\ell} \mid \mathcal{H}_{q,\ell}) &\leq \sum_{u\in\mathcal{U}} \mathbb{P}_\mu\big(\theta_{u,T_q+\ell} \notin [1/t, 1-1/t] \mid \mathcal{H}_{q,\ell}\big) \\ &\quad + \sum_{\substack{u,v\in\mathcal{U}\\ u<v}} \mathbb{E}_\mu\big[\mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{v,T_q+\ell}| \leq \delta_t \mid \mathcal{H}_{q,\ell}, \theta_{v,T_q+\ell}\big)\big|\, \mathcal{H}_{q,\ell}\big] \\ &\quad + \sum_{u\in\mathcal{U}} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{a^\star(\mu),T_q+\ell}| \leq \delta_t \mid \mathcal{H}_{q,\ell}\big) \\ &\leq \sum_{u\in\mathcal{U}} \Big(\mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - 1/t| \leq \delta_t \mid \mathcal{H}_{q,\ell}\big) + \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - (1-1/t)| \leq \delta_t \mid \mathcal{H}_{q,\ell}\big)\Big) \\ &\quad + \sum_{\substack{u,v\in\mathcal{U}\\ u<v}} \mathbb{E}_\mu\big[\mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{v,T_q+\ell}| \leq \delta_t \mid \mathcal{H}_{q,\ell}, \theta_{v,T_q+\ell}\big)\big|\, \mathcal{H}_{q,\ell}\big] \\ &\quad + \sum_{u\in\mathcal{U}} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{a^\star(\mu),T_q+\ell}| \leq \delta_t \mid \mathcal{H}_{q,\ell}\big) \\ &\leq 2\sum_{u\in\mathcal{U}} \sup_{x\in(0,1)} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - x| \leq \delta_t \mid \mathcal{H}_{q,\ell}\big) \\ &\quad + \sum_{\substack{u,v\in\mathcal{U}\\ u<v}} \sup_{x\in(0,1)} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - x| \leq \delta_t \mid \mathcal{H}_{q,\ell}\big) \\ &\quad + \sum_{u\in\mathcal{U}} \sup_{x\in(0,1)} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - x| \leq \delta_t \mid \mathcal{H}_{q,\ell}\big) \end{aligned}$$

$$
\begin{aligned}
&\overset{(8)}{\le} 2|\mathcal{U}|\,\frac{C_0(1+A)}{(\log t)^3} \\
&\quad + \binom{|\mathcal{U}|}{2}\frac{C_0(1+A)}{(\log t)^3} \\
&\quad + |\mathcal{U}|\,\frac{C_0(1+A)}{(\log t)^3} \\
&= \Big(3|\mathcal{U}| + \binom{|\mathcal{U}|}{2}\Big)\frac{C_0(1+A)}{(\log t)^3} \\
&\le \Big(2K + \binom{K}{2} + K\Big)\frac{C_0(1+A)}{(\log t)^3} = \frac{\big(3K + \binom{K}{2}\big)C_0(1+A)}{(\log t)^3},
\end{aligned}
$$

Here the factor $2|\mathcal{U}|$ comes from the two boundary events for each $u \in \mathcal{U}$, the factor $\binom{|\mathcal{U}|}{2}$ comes from the pairs inside $\mathcal{U}$ after conditioning on one arm draw and applying the same supremum bound, and the final factor $|\mathcal{U}|$ comes from the mixed events involving $\theta_{a^\star(\mu),T_q+\ell}$. After enlarging the lower threshold on t once more,

$$
\mathbb{P}_\mu(G_{T_q+\ell} \mid \mathcal{H}_{q,\ell}) \ge \frac{1}{2}. \tag{G}
$$

On

$$
G_{T_q+\ell} \cap \Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\},
$$

the hypotheses of Lemma 12 are satisfied with $U = \mathcal{U}$: by the definition of $G_{T_q+\ell}$, every unsaturated arm draw lies in the interior, these draws are pairwise δ_t-separated, and they are δ_t-separated from $\theta_{a^\star(\mu),T_q+\ell}$; on $\mathrm{Good}_{T_q+\ell}$ every saturated arm draw lies in the interior; and by (13), every saturated arm $k \neq a^\star(\mu)$ is δ_t-separated from $\theta_{a^\star(\mu),T_q+\ell}$. Therefore, by Lemma 12,

$$
\sum_{u\in\mathcal{U}} w_u^\star(\theta_{T_q+\ell}) \ge \frac{1}{2K(\log t)^9}. \qquad \text{a.s. on } G_{T_q+\ell} \cap \Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\}.
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}_\mu\big(I_{T_q+\ell} \in \mathcal{U},\ \Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,\ell}\big) &= \mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\mathbb{P}_\mu\big(I_{T_q+\ell} \in \mathcal{U} \mid \mathcal{H}_{q,\ell}\big) \\
&= \mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\mathbb{E}_\mu\left[\sum_{u\in\mathcal{U}} w_u^\star(\theta_{T_q+\ell}) \,\middle|\, \mathcal{H}_{q,\ell}\right] \\
&\ge \frac{\mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{S}_{T_q},\ \mathrm{Good}_{T_q+\ell}\}}{2K(\log t)^9}\mathbb{P}_\mu(G_{T_q+\ell} \mid \mathcal{H}_{q,\ell}) \\
&\ge \frac{\mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{S}_{T_q},\ \mathrm{Good}_{T_q+\ell}\}}{4K(\log t)^9}.
\end{aligned}
$$

Here the first equality uses the $\mathcal{H}_{q,\ell}$-measurability of $\Xi_{\ell-1} \cap \{a^\star(\mu) \in \mathcal{S}_{T_q}\}$, the second uses the sampling rule, and the inequalities use the pointwise bound above together with (G). Equivalently,

$$
\mathbb{P}_\mu\big(I_{T_q+\ell} \in \mathcal{S}_{T_q},\ \Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,\ell}\big) \le \mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\left(1 - \frac{\mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}{4K(\log t)^9}\right). \tag{14}
$$

(b)-2. Iterating the one-step bound along the block. We now iterate (14) along the block. For $1 \le r \le L'$, write $\mathcal{H}_{q,r}$ for the same sigma-field from Lemma 11, and define

$$
p_t := 1 - \frac{1}{4K(\log t)^9}.
$$

Then

$$
1 - \frac{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}{4K(\log t)^9} = {p_t}^{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}.
$$

Using the identity $\mathbb{P}(A \mid \mathcal{G}) = \mathbb{E}[\mathbf{1}\{A\} \mid \mathcal{G}]$ together with (10) and the tower property, we obtain

$$
\begin{aligned}
&\mathbb{P}_\mu\big(\Xi_r,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{F}_{T_q}\big) \\
&\quad = \mathbb{E}_\mu\big[\mathbf{1}\{\Xi_r,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\big|\,\mathcal{F}_{T_q}\big] \\
&\quad = \mathbb{E}_\mu\big\{\mathbb{E}_\mu\big[\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\mathbf{1}\{I_{T_q+r} \in \mathcal{S}_{T_q}\}\big|\,\mathcal{H}_{q,r}\big]\big|\,\mathcal{F}_{T_q}\big\} \\
&\quad = \mathbb{E}_\mu\big[\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\mathbb{P}_\mu(I_{T_q+r} \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,r})\big|\,\mathcal{F}_{T_q}\big].
\end{aligned} \tag{15}
$$

Here the second equality uses (10) and the tower property, and the third uses the $\mathcal{H}_{q,r}$-measurability of $\Xi_{r-1} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}$. Equation (14), with r in place of ℓ, gives

$$\begin{aligned}&\mathbb{P}_\mu\big(I_{T_q+r} \in \mathcal{S}_{T_q},\ \Xi_{r-1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,r}\big)\\ &\quad\le \mathbf{1}\{\Xi_{r-1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\left(1 - \frac{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}{4K(\log t)^9}\right) \qquad \text{a.s.}\end{aligned}$$

Multiplying both sides by the $\mathcal{H}_{q,r}$-measurable indicator $\mathbf{1}\{K \le T_q < \tau_{i,j+1}\}$, we obtain

$$\begin{aligned}&\mathbb{P}_\mu\big(I_{T_q+r} \in \mathcal{S}_{T_q},\ \Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,r}\big)\\ &\quad\le \mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\left(1 - \frac{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}{4K(\log t)^9}\right) \qquad \text{a.s.}\end{aligned} \tag{16}$$

Since $\Xi_{r-1} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\} \in \mathcal{H}_{q,r}$, the conditional probability on the left is

$$\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\mathbb{P}_\mu(I_{T_q+r} \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,r}).$$

Taking $\mathbb{E}_\mu[\cdot \mid \mathcal{F}_{T_q}]$ of (16) gives

$$\begin{aligned}&\mathbb{E}_\mu\big[\mathbb{P}_\mu\big(I_{T_q+r} \in \mathcal{S}_{T_q},\ \Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,r}\big)\big|\,\mathcal{F}_{T_q}\big]\\ &\quad\le \mathbb{E}_\mu\left[\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\left(1 - \frac{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}{4K(\log t)^9}\right)\middle|\,\mathcal{F}_{T_q}\right].\end{aligned}$$

By (15), the left-hand side is

$$\mathbb{P}_\mu\big(\Xi_r,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{F}_{T_q}\big),$$

so

$$\begin{aligned}&\mathbb{P}_\mu\big(\Xi_r,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{F}_{T_q}\big)\\ &\quad\le \mathbb{E}_\mu\left[\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\left(1 - \frac{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}{4K(\log t)^9}\right)\middle|\,\mathcal{F}_{T_q}\right]\\ &\quad= \mathbb{E}_\mu\left[\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}p_t{}^{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}\middle|\,\mathcal{F}_{T_q}\right].\end{aligned} \tag{17}$$

(b)-3. From the inverse-factor recurrence to exponential decay. The difficulty is that the number and locations of the good rounds inside the block are random, so (17) cannot be turned directly into a deterministic power of p_t before reaching the end of the block. We therefore first encode the event $\{\mathrm{Bad} \le L'/2\}$ through the inverse factor below, and only then extract the deterministic power $p_t{}^{\lceil L'/2\rceil}$. On $\{\mathrm{Bad} \le L'/2\}$,

$$\sum_{r=1}^{L'} \mathbf{1}\{\mathrm{Good}_{T_q+r}\} = L' - \mathrm{Bad} \ge \lceil L'/2\rceil,$$

so

$$\begin{aligned}&\mathbf{1}\{\Xi_{L'},\ \mathrm{Bad} \le L'/2,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\\ &\quad\le \mathbf{1}\{\Xi_{L'},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}p_t{}^{-\sum_{r=1}^{L'}\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}p_t{}^{\lceil L'/2\rceil}.\end{aligned}$$

Taking conditional expectations with respect to $\mathcal{F}_{T_q}$ gives

$$\begin{aligned}&\mathbb{P}_\mu\big(\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{F}_{T_q}\big)\\ &\quad\le p_t{}^{\lceil L'/2\rceil}\mathbb{E}_\mu\left[\mathbf{1}\{\Xi_{L'} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\}p_t{}^{-\sum_{\ell=1}^{L'}\mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}\middle|\,\mathcal{F}_{T_q}\right].\end{aligned} \tag{18}$$

The inverse factor $p_t{}^{-\sum_{\ell=1}^{r}\mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}$ is $\mathcal{H}_{q,r}$-measurable. Multiplying (16) by this factor, using (10), and applying the tower property gives

$$\begin{aligned}&\mathbb{E}_\mu\left[\mathbf{1}\{\Xi_r,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}p_t{}^{-\sum_{\ell=1}^{r}\mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}\middle|\,\mathcal{F}_{T_q}\right]\\ &\quad\le \mathbb{E}_\mu\left[\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}p_t{}^{-\sum_{\ell=1}^{r-1}\mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}\middle|\,\mathcal{F}_{T_q}\right].\end{aligned}$$

Iterating over $r = 1, \dots, L'$ yields

$$\begin{aligned}\mathbb{E}_\mu\Big[\mathbf{1}\{\Xi_{L'} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}\} p_t^{-\sum_{\ell=1}^{L'} \mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}} \Big| \mathcal{F}_{T_q}\Big] \\ \le \mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\}.\end{aligned} \tag{19}$$

Combining (18) and (19), we obtain

$$\begin{aligned}&\mathbb{P}_\mu\big(\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q} \mid \mathcal{F}_{T_q}\big) \\ &\quad \le \mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\} p_t^{\lceil L'/2 \rceil} \\ &\quad \le \mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{S}_{T_q}\} \exp\Big(-\frac{\lceil L'/2 \rceil}{4K(\log t)^9}\Big).\end{aligned}$$

and taking expectation gives

$$(b) \le \exp\Big(-\frac{\lceil L'/2 \rceil}{4K(\log t)^9}\Big).$$

Since $L' \ge (16KA^2)^{-1} t/(\log t)^2$, this is at most

$$\exp\Big(-\frac{t}{128K^2A^2(\log t)^{11}}\Big)$$

for all large enough t.

$$(b) \le \exp\Big(-\frac{t}{128K^2A^2(\log t)^{11}}\Big).$$

4. Bounding term (c). We now consider the event in term (c):

$$\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\} \cap \{a^\star(\mu) \in \mathcal{U}\}.$$

We proceed in the three stages used in the proof sketch: first derive and iterate the one-step conditional escape bound using the $\mathcal{F}_{T_q}$-measurable quantity Φ_q; next establish a lower bound on Φ_q on the event $\mathcal{M}_q$ and combine it with the block-length estimate; and finally control the contribution from $\mathcal{M}_q^c$ when $a^\star(\mu) \in \mathcal{U}$. Here

$$\Phi_q := \mathbb{P}_\mu\Big(\theta_{a^\star(\mu),T_q+1} \ge \mu^-_{a^\star(\mu)} \ \Big|\ \mathcal{F}_{T_q}\Big).$$

This is the conditional posterior probability that the draw of the true best arm is at least $\mu^-_{a^\star(\mu)}$; it is needed only in the case $a^\star(\mu) \in \mathcal{U}$. In other words, to bound

$$\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\} \cap \{a^\star(\mu) \in \mathcal{U}\},$$

we first study, for each ℓ, the local event

$$\begin{aligned}&\Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell} \cap \{a^\star(\mu) \in \mathcal{U}\} \\ &\quad = \{K \le T_q < \tau_{i,j+1}\} \cap \bigcap_{m=1}^{\ell-1} \{I_{T_q+m} \in \mathcal{S}_{T_q}\} \cap \bigcap_{k \in \mathcal{S}_{T_q}} \{|\theta_{k,T_q+\ell} - \mu_k| \le r_\mu\} \cap \{a^\star(\mu) \in \mathcal{U}\},\end{aligned}$$

and then combine those one-step bounds across the block.

(c)-1. One-step conditional escape bound.

Fix $\ell \in \{1, \dots, L'\}$ and let $\mathcal{H}_{q,\ell}$ be the sigma-field from Lemma 11. Then $\Xi_{\ell-1}$ and $\mathrm{Good}_{T_q+\ell}$ are $\mathcal{H}_{q,\ell}$-measurable. On $\Xi_{\ell-1}$, no arm in $\mathcal{U}$ is sampled during $\{T_q+1, \dots, T_q+\ell-1\}$, so the posterior parameters of all unsaturated arms remain equal to their values at time T_q. Hence, conditional on $\mathcal{H}_{q,\ell}$, the current posterior draws of the unsaturated arms at round $T_q+\ell$ are independent and satisfy the conditioned one-dimensional and pairwise anti-concentration bounds from Lemma 11. In particular, Φ_q is already $\mathcal{H}_{q,\ell}$-measurable, and on $\Xi_{\ell-1} \cap \{a^\star(\mu) \in \mathcal{U}\}$ we have

$$\mathbb{P}_\mu\Big(\theta_{a^\star(\mu),T_q+\ell} \ge \mu^-_{a^\star(\mu)} \ \Big|\ \mathcal{H}_{q,\ell}\Big) = \Phi_q.$$

To obtain the one-step escape bound for term (c), we now work under the conjunction

$$\Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell} \cap \{a^\star(\mu) \in \mathcal{U}\} \cap \Big\{\theta_{a^\star(\mu),T_q+\ell} \ge \mu^-_{a^\star(\mu)}\Big\}.$$

On this event, every saturated arm $k \in \mathcal{S}_{T_q}$ satisfies

$$\theta_{a^\star(\mu),T_q+\ell} - \theta_{k,T_q+\ell} \geq \mu^-_{a^\star(\mu)} - (\mu_k + r_\mu) \geq \frac{3\Delta_{\min}}{4} - r_\mu \geq \frac{5\Delta_{\min}}{8} > \delta_t. \tag{20}$$

Moreover, on $\mathrm{Good}_{T_q+\ell}$ the posterior draw of every saturated arm lies in $[1/t, 1-1/t]$ for all large enough t. We now define an auxiliary event $G'_{T_q+\ell}$ on which every unsaturated arm other than $a^\star(\mu)$ lies in $[1/t, 1-1/t]$, the draws of distinct such arms are more than δ_t apart, and every such arm draw is more than δ_t away from $\theta_{a^\star(\mu),T_q+\ell}$:

$$\begin{aligned} G'_{T_q+\ell} :=& \Big\{\theta_{u,T_q+\ell} \in [1/t, 1-1/t] \text{ for all } u \in \mathcal{U} \setminus \{a^\star(\mu)\}\Big\} \\ &\cap \bigcap_{\substack{u,v \in \mathcal{U}\setminus\{a^\star(\mu)\} \\ u \neq v}} \{|\theta_{u,T_q+\ell} - \theta_{v,T_q+\ell}| > \delta_t\} \cap \bigcap_{u \in \mathcal{U}\setminus\{a^\star(\mu)\}} \{|\theta_{u,T_q+\ell} - \theta_{a^\star(\mu),T_q+\ell}| > \delta_t\}. \end{aligned}$$

For brevity, write

$$\widehat{\mathcal{H}}_{q,\ell} := \sigma\big(\mathcal{H}_{q,\ell}, \theta_{a^\star(\mu),T_q+\ell}\big).$$

The event $(G'_{T_q+\ell})^c$ can occur only in three ways: some unsaturated arm other than $a^\star(\mu)$ leaves the interval $[1/t, 1-1/t]$, two such arm draws come within distance δ_t, or one of them comes within distance δ_t of $\theta_{a^\star(\mu),T_q+\ell}$. We bound these three families separately. By the conditioned variant of (8) from Lemma 11 and the corresponding pairwise anti-concentration bound obtained from the same lemma by conditional independence, and since $1/t \leq \delta_t$, we have

$$\{\theta_{u,T_q+\ell} \notin [1/t, 1-1/t]\} \subseteq \{|\theta_{u,T_q+\ell} - 1/t| \leq \delta_t\} \cup \{|\theta_{u,T_q+\ell} - (1-1/t)| \leq \delta_t\}$$

for every $u \in \mathcal{U} \setminus \{a^\star(\mu)\}$. Therefore a union bound gives

$$\begin{aligned} \mathbb{P}_\mu\big((G'_{T_q+\ell})^c \mid \widehat{\mathcal{H}}_{q,\ell}\big) \leq & \sum_{u \in \mathcal{U}\setminus\{a^\star(\mu)\}} \mathbb{P}_\mu\big(\theta_{u,T_q+\ell} \notin [1/t, 1-1/t] \mid \widehat{\mathcal{H}}_{q,\ell}\big) \\ &+ \sum_{\substack{u,v \in \mathcal{U}\setminus\{a^\star(\mu)\} \\ u<v}} \mathbb{E}_\mu\Big[\mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{v,T_q+\ell}| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}, \theta_{v,T_q+\ell}\big)\Big| \widehat{\mathcal{H}}_{q,\ell}\Big] \\ &+ \sum_{u \in \mathcal{U}\setminus\{a^\star(\mu)\}} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{a^\star(\mu),T_q+\ell}| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}\big) \\ \leq & \sum_{u \in \mathcal{U}\setminus\{a^\star(\mu)\}} \Big(\mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - 1/t| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}\big) + \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - (1-1/t)| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}\big)\Big) \\ &+ \sum_{\substack{u,v \in \mathcal{U}\setminus\{a^\star(\mu)\} \\ u<v}} \mathbb{E}_\mu\Big[\mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{v,T_q+\ell}| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}, \theta_{v,T_q+\ell}\big)\Big| \widehat{\mathcal{H}}_{q,\ell}\Big] \\ &+ \sum_{u \in \mathcal{U}\setminus\{a^\star(\mu)\}} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - \theta_{a^\star(\mu),T_q+\ell}| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}\big) \\ \leq & 2 \sum_{u \in \mathcal{U}\setminus\{a^\star(\mu)\}} \sup_{x \in (0,1)} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - x| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}\big) \\ &+ \sum_{\substack{u,v \in \mathcal{U}\setminus\{a^\star(\mu)\} \\ u<v}} \sup_{x \in (0,1)} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - x| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}\big) \\ &+ \sum_{u \in \mathcal{U}\setminus\{a^\star(\mu)\}} \sup_{x \in (0,1)} \mathbb{P}_\mu\big(|\theta_{u,T_q+\ell} - x| \leq \delta_t \mid \widehat{\mathcal{H}}_{q,\ell}\big) \end{aligned}$$

$$
\begin{aligned}
&\overset{(8)}{\leq} 2|\mathcal{U}|\,\frac{C_0(1+A)}{(\log t)^3} \\
&\quad + \binom{|\mathcal{U}|}{2}\frac{C_0(1+A)}{(\log t)^3} \\
&\quad + |\mathcal{U}|\,\frac{C_0(1+A)}{(\log t)^3} \\
&= \Big(2|\mathcal{U}| + \binom{|\mathcal{U}|}{2} + |\mathcal{U}|\Big)\frac{C_0(1+A)}{(\log t)^3} \\
&\leq \Big(2K + \binom{K}{2} + K\Big)\frac{C_0(1+A)}{(\log t)^3} \\
&= \frac{\big(3K + \binom{K}{2}\big)C_0(1+A)}{(\log t)^3}.
\end{aligned}
$$

Here the factor $2|\mathcal{U}|$ comes from the two boundary events per arm, the factor $\binom{|\mathcal{U}|}{2}$ from unordered pairs of distinct unsaturated arms, and the final factor $|\mathcal{U}|$ from the mixed events involving $\theta_{a^\star(\mu),T_q+\ell}$. After enlarging the lower threshold on t once more,

$$
\mathbb{P}_\mu\big(G'_{T_q+\ell} \mid \widehat{\mathcal{H}}_{q,\ell}\big) \geq \frac{1}{2}. \tag{G'}
$$

On

$$
G'_{T_q+\ell} \cap \Big\{\theta_{a^\star(\mu),T_q+\ell} \geq \mu^-_{a^\star(\mu)}\Big\} \cap \Xi_{\ell-1} \cap \mathrm{Good}_{T_q+\ell} \cap \{a^\star(\mu) \in \mathcal{U}\}, \tag{21}
$$

set $x := \theta_{T_q+\ell}$, and let a and b be a best arm and a second-best arm of x, respectively. By the definition of $G'_{T_q+\ell}$, every arm in $\mathcal{U} \setminus \{a^\star(\mu)\}$ lies in $[1/t, 1-1/t]$, these draws are pairwise δ_t-separated, and they are δ_t-separated from $x_{a^\star(\mu)}$; on $\mathrm{Good}_{T_q+\ell}$ every saturated arm draw lies in $[1/t, 1-1/t]$; and by (20), every saturated arm lies more than δ_t below $x_{a^\star(\mu)}$. Indeed, either $a = a^\star(\mu)$, in which case every other arm lies at least δ_t below x_a by $G'_{T_q+\ell}$ and (20), or else $a \in \mathcal{U} \setminus \{a^\star(\mu)\}$, in which case all saturated arms lie more than δ_t below $x_{a^\star(\mu)} < x_a$, while every other unsaturated arm is δ_t-separated from x_a by $G'_{T_q+\ell}$. Consequently the best and second-best values of x are separated by at least δ_t, so the lower-bound part of the proof of Lemma 12 gives

$$
C^\star(x) \geq \frac{\delta_t^2}{2K}.
$$

We do not apply Lemma 12 itself here. Instead, we reuse the one-alternative upper-bound part of its proof with the specific choice $u = a^\star(\mu)$ to force a lower bound on $w^\star_{a^\star(\mu)}(x)$. To upper bound $C^\star(x)$, we again use the proof of Lemma 12. If $a = a^\star(\mu)$, then $b \neq a^\star(\mu)$ and hence $x_b \in [1/t, 1-1/t]$; applying the case $u = a$ in that proof yields

$$
C^\star(x) \leq w^\star_{a^\star(\mu)}(x)\, d(x_{a^\star(\mu)}, x_b) \leq w^\star_{a^\star(\mu)}(x) \log t.
$$

If instead $a \neq a^\star(\mu)$, then $a \in \mathcal{U} \setminus \{a^\star(\mu)\}$ and therefore $x_a \in [1/t, 1-1/t]$; applying the case $u = a^\star(\mu) \neq a$ in that proof gives

$$
C^\star(x) \leq w^\star_{a^\star(\mu)}(x)\, d(x_{a^\star(\mu)}, x_a) \leq w^\star_{a^\star(\mu)}(x) \log t.
$$

Combining the lower bound on $C^\star(x)$ with the upper bound in the case $a = a^\star(\mu)$ gives

$$
w^\star_{a^\star(\mu)}(x) \geq \frac{C^\star(x)}{\log t} \geq \frac{\delta_t^2}{2K \log t} = \frac{1}{2K(\log t)^9}.
$$

Combining the same lower bound with the upper bound in the case $a \neq a^\star(\mu)$ gives exactly the same conclusion. Hence

$$
w^\star_{a^\star(\mu)}(x) \geq \frac{1}{2K(\log t)^9}.
$$

Since $a^\star(\mu) \in \mathcal{U}$ on the event in (21), it follows that

$$
\sum_{u \in \mathcal{U}} w^\star_u(\theta_{T_q+\ell}) \geq w^\star_{a^\star(\mu)}(\theta_{T_q+\ell}) \geq \frac{1}{2K(\log t)^9}. \qquad \text{a.s. on the event in (21).} \tag{22}
$$

As in the analysis of term (b),

$$\mathbb{P}_\mu\big(I_{T_q+\ell} \in \mathcal{U} \mid \mathcal{F}_{T_q+\ell-1}, \theta_{T_q+\ell}\big) = \sum_{u\in\mathcal{U}} w_u^\star(\theta_{T_q+\ell}).$$

Taking conditional expectations with respect to $\mathcal{H}_{q,\ell}$ therefore yields

$$\begin{aligned}
&\mathbb{P}_\mu\big(I_{T_q+\ell} \in \mathcal{U},\ \Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{H}_{q,\ell}\big) \\
&\quad= \mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U}\}\mathbb{P}_\mu\big(I_{T_q+\ell} \in \mathcal{U} \mid \mathcal{H}_{q,\ell}\big) \\
&\quad= \mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U}\}\mathbb{E}_\mu\left[\sum_{u\in\mathcal{U}} w_u^\star(\theta_{T_q+\ell}) \,\middle|\, \mathcal{H}_{q,\ell}\right] \\
&\quad\geq \frac{\mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U},\ \mathrm{Good}_{T_q+\ell}\}}{2K(\log t)^9}\mathbb{E}_\mu\Big[\mathbf{1}\Big\{\theta_{a^\star(\mu),T_q+\ell} \geq \mu^-_{a^\star(\mu)}\Big\} \\
&\qquad\qquad \times \mathbb{P}_\mu(G'_{T_q+\ell} \mid \mathcal{H}_{q,\ell}, \theta_{a^\star(\mu),T_q+\ell})\Big|\, \mathcal{H}_{q,\ell}\Big] \\
&\quad\geq \frac{\Phi_q}{4K(\log t)^9}\mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U},\ \mathrm{Good}_{T_q+\ell}\} \\
&\quad\geq \frac{\Phi_q}{8K(\log t)^9}\mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U}\}\mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}.
\end{aligned}$$

Here the first equality uses the $\mathcal{H}_{q,\ell}$-measurability of $\Xi_{\ell-1} \cap \{a^\star(\mu) \in \mathcal{U}\}$, the second uses the sampling rule, the first inequality uses the pointwise bound in (22) together with the $\mathcal{H}_{q,\ell}$-measurability of $\mathrm{Good}_{T_q+\ell}$, the second uses (G') and the definition of Φ_q, and the last line only weakens the constant from $1/4$ to $1/8$ for later convenience. Equivalently,

$$\mathbb{P}_\mu\big(I_{T_q+\ell} \in \mathcal{S}_{T_q},\ \Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{H}_{q,\ell}\big) \leq \mathbf{1}\{\Xi_{\ell-1},\ a^\star(\mu) \in \mathcal{U}\}\left(1 - \frac{\Phi_q\,\mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}{8K(\log t)^9}\right). \tag{23}$$

Iteration along the block. We now iterate (23) along the block. For $1 \leq r \leq L'$, define

$$q^U := 1 - \frac{\Phi_q}{8K(\log t)^9}.$$

Then

$$1 - \frac{\Phi_q\,\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}{8K(\log t)^9} = (q^U)^{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}.$$

Since Φ_q is $\mathcal{F}_{T_q}$-measurable, the random coefficient q^U is already $\mathcal{H}_{q,r}$-measurable; the same is true for $\{K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}$. Using the identity $\mathbb{P}(A \mid \mathcal{G}) = \mathbb{E}[\mathbf{1}\{A\} \mid \mathcal{G}]$ together with (10) and the tower property, we obtain

$$\begin{aligned}
&\mathbb{P}_\mu\big(\Xi_r,\ K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{F}_{T_q}\big) \\
&\quad= \mathbb{E}_\mu\big[\mathbf{1}\{\Xi_r,\ K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\big|\, \mathcal{F}_{T_q}\big] \\
&\quad= \mathbb{E}_\mu\big\{\mathbb{E}_\mu\big[\mathbf{1}\{\Xi_{r-1},\ K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\mathbf{1}\{I_{T_q+r} \in \mathcal{S}_{T_q}\}\big|\, \mathcal{H}_{q,r}\big]\big|\, \mathcal{F}_{T_q}\big\} \\
&\quad= \mathbb{E}_\mu\big[\mathbf{1}\{\Xi_{r-1},\ K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\mathbb{P}_\mu(I_{T_q+r} \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,r})\big|\, \mathcal{F}_{T_q}\big].
\end{aligned} \tag{24}$$

Here the second equality uses (10) and the tower property, and the third uses the $\mathcal{H}_{q,r}$-measurability of $\Xi_{r-1} \cap \{K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}$. Equation (23), with r in place of ℓ, gives

$$\begin{aligned}
&\mathbb{P}_\mu\big(I_{T_q+r} \in \mathcal{S}_{T_q},\ \Xi_{r-1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{H}_{q,r}\big) \\
&\qquad\leq \mathbf{1}\{\Xi_{r-1},\ a^\star(\mu) \in \mathcal{U}\}(q^U)^{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}} \qquad \text{a.s.}
\end{aligned}$$

Multiplying both sides by the $\mathcal{H}_{q,r}$-measurable indicator $\mathbf{1}\{K \leq T_q < \tau_{i,j+1}\}$, we obtain

$$\begin{aligned}
&\mathbb{P}_\mu\big(I_{T_q+r} \in \mathcal{S}_{T_q},\ \Xi_{r-1},\ K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{H}_{q,r}\big) \\
&\qquad\leq \mathbf{1}\{\Xi_{r-1},\ K \leq T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}(q^U)^{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}} \qquad \text{a.s.}
\end{aligned} \tag{25}$$

Since $\Xi_{r-1} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} \in \mathcal{H}_{q,r}$, the conditional probability on the left is

$$\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\mathbb{P}_\mu(I_{T_q+r} \in \mathcal{S}_{T_q} \mid \mathcal{H}_{q,r}).$$

Taking $\mathbb{E}_\mu[\cdot \mid \mathcal{F}_{T_q}]$ of (25) gives

$$\begin{aligned}&\mathbb{E}_\mu\big[\mathbb{P}_\mu\big(I_{T_q+r} \in \mathcal{S}_{T_q},\ \Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{H}_{q,r}\big)\big|\,\mathcal{F}_{T_q}\big]\\ &\quad\le \mathbb{E}_\mu\Big[\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}(q^U)^{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}\Big|\,\mathcal{F}_{T_q}\Big].\end{aligned}$$

By (24), the left-hand side is $\mathbb{P}_\mu\big(\Xi_r,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{F}_{T_q}\big)$, so

$$\begin{aligned}\mathbb{P}_\mu\big(\Xi_r,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{F}_{T_q}\big) \le \mathbb{E}_\mu\Big[&\mathbf{1}\{\Xi_{r-1},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\\ &\times (q^U)^{\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}\ \Big|\ \mathcal{F}_{T_q}\Big].\end{aligned} \tag{26}$$

Using the number of good rounds. As in the analysis of term (b), the number and locations of the good rounds inside the block are random, so (26) cannot be turned directly into a deterministic power before reaching the end of the block. Here the coefficient Φ_q is also random, but it is already fixed at time T_q, so the same is true for q^U. On $\{\mathrm{Bad} \le L'/2\}$,

$$\sum_{r=1}^{L'} \mathbf{1}\{\mathrm{Good}_{T_q+r}\} = L' - \mathrm{Bad} \ge \lceil L'/2 \rceil,$$

so

$$\begin{aligned}&\mathbf{1}\{\Xi_{L'},\ \mathrm{Bad} \le L'/2,\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\\ &\quad\le \mathbf{1}\{\Xi_{L'},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}(q^U)^{-\sum_{\ell=1}^{L'} \mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}(q^U)^{\lceil L'/2 \rceil}.\end{aligned}$$

Taking conditional expectations with respect to $\mathcal{F}_{T_q}$ gives

$$\begin{aligned}&\mathbb{P}_\mu\big(\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{F}_{T_q}\big)\\ &\quad\le (q^U)^{\lceil L'/2 \rceil}\mathbb{E}_\mu\Big[\mathbf{1}\{\Xi_{L'} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\}(q^U)^{-\sum_{\ell=1}^{L'} \mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}\Big|\,\mathcal{F}_{T_q}\Big].\end{aligned} \tag{27}$$

The inverse factor $(q^U)^{-\sum_{\ell=1}^{r} \mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}$ is $\mathcal{H}_{q,r}$-measurable. Multiplying (25) by this factor, using (10), and applying the tower property gives

$$\begin{aligned}&\mathbb{E}_\mu\Big[\mathbf{1}\{\Xi_r \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\}\,(q^U)^{-\sum_{\ell=1}^{r} \mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}\Big|\,\mathcal{F}_{T_q}\Big]\\ &\quad\le \mathbb{E}_\mu\Big[\mathbf{1}\{\Xi_{r-1} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\}\,(q^U)^{-\sum_{\ell=1}^{r-1} \mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}\Big|\,\mathcal{F}_{T_q}\Big].\end{aligned}$$

In other words, multiplication by the inverse factor $(q^U)^{-\mathbf{1}\{\mathrm{Good}_{T_q+r}\}}$ compensates for the one-step contraction at round r, so the conditional expectation of the resulting quantity does not increase with r. Iterating over $r = 1, \dots, L'$ yields

$$\begin{aligned}&\mathbb{E}_\mu\Big[\mathbf{1}\{\Xi_{L'} \cap \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\}\,(q^U)^{-\sum_{\ell=1}^{L'} \mathbf{1}\{\mathrm{Good}_{T_q+\ell}\}}\Big|\,\mathcal{F}_{T_q}\Big]\\ &\quad\le \mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}.\end{aligned} \tag{28}$$

Combining (27) and (28), we obtain

$$\begin{aligned}&\mathbb{P}_\mu\big(\Xi_{L'} \cap \{\mathrm{Bad} \le L'/2\},\ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U} \mid \mathcal{F}_{T_q}\big)\\ &\quad\le \mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}(q^U)^{\lceil L'/2 \rceil}\\ &\quad\le \mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\exp\Big(-\frac{\Phi_q\,\lceil L'/2 \rceil}{8K(\log t)^9}\Big).\end{aligned}$$

Taking expectation gives

$$(c) \le \mathbb{E}_\mu\left[\exp\Big(-\frac{\Phi_q\,\lceil L'/2 \rceil}{8K(\log t)^9}\Big)\mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\right].$$

(c)-2. Lower bound on Φ_q on $\mathcal{M}_q$. This is the only place in the analysis of term (c) where an unusually small value of Φ_q could spoil the escape estimate. We therefore split according to whether the KL quantity at time T_q is small enough to yield a usable lower bound on Φ_q. Let

$$\mathcal{M}_q := \Big\{ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \notin \mathcal{S}_{T_q},\ N_{a^\star(\mu)}(T_q)\,\mathrm{kl}_-\Big(\frac{S_{a^\star(\mu)}(T_q)}{N_{a^\star(\mu)}(T_q)}, \mu^-_{a^\star(\mu)}\Big) \le \frac{1+R_\mu}{2R_\mu}\log t \Big\}.$$

Since

$$\mathcal{M}_q \subseteq \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \notin \mathcal{S}_{T_q}\} = \{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\},$$

we have

$$\mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} = \mathbf{1}\{\mathcal{M}_q\} + \mathbf{1}\{\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} \cap \mathcal{M}_q^c\}.$$

Therefore

$$\begin{aligned}
(c) &\le \mathbb{E}_\mu\Big[\exp\Big(-\frac{\Phi_q\lceil L'/2\rceil}{8K(\log t)^9}\Big)\mathbf{1}\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\}\Big] \\
&= \mathbb{E}_\mu\Big[\exp\Big(-\frac{\Phi_q\lceil L'/2\rceil}{8K(\log t)^9}\Big)\mathbf{1}\{\mathcal{M}_q\}\Big] \\
&\quad + \mathbb{E}_\mu\Big[\exp\Big(-\frac{\Phi_q\lceil L'/2\rceil}{8K(\log t)^9}\Big)\mathbf{1}\{\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} \cap \mathcal{M}_q^c\}\Big] \\
&\le \mathbb{E}_\mu\Big[\exp\Big(-\frac{\Phi_q\lceil L'/2\rceil}{8K(\log t)^9}\Big)\mathbf{1}\{\mathcal{M}_q\}\Big] \\
&\quad + \mathbb{P}_\mu\big(\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} \cap \mathcal{M}_q^c\big),
\end{aligned}$$

where the last inequality uses

$$\exp\Big(-\frac{\Phi_q\lceil L'/2\rceil}{8K(\log t)^9}\Big) \le 1$$

because $\Phi_q \ge 0$.

On $\mathcal{M}_q$, Lemma 14 gives

$$\Phi_q \ge c\,(A\log t + 1)^{-1/2} t^{-(1+R_\mu)/(2R_\mu)}.$$

Since $L' \ge (16KA^2)^{-1}t/(\log t)^2$ for all large enough t, it follows that

$$\frac{\Phi_q}{8K(\log t)^9}\Big\lceil\frac{L'}{2}\Big\rceil \ge c\,\frac{t^{(R_\mu-1)/(2R_\mu)}}{(\log t)^{14}} \tag{29}$$

there. Hence

$$\mathbb{E}_\mu\Big[\exp\Big(-\frac{\Phi_q\lceil L'/2\rceil}{8K(\log t)^9}\Big)\mathbf{1}\{\mathcal{M}_q\}\Big] \le \exp\Big(-c\,\frac{t^{(R_\mu-1)/(2R_\mu)}}{(\log t)^{14}}\Big).$$

(c)-3. Bounding the contribution from $\mathcal{M}_q^c$. It therefore remains to control the probability

$$\mathbb{P}_\mu\big(\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} \cap \mathcal{M}_q^c\big).$$

Because $a^\star(\mu) \notin \mathcal{S}_{T_q}$ implies $1 \le N_{a^\star(\mu)}(T_q) \le \lfloor A\log t\rfloor$, we may sum over $n \in \{1,\dots,\lfloor A\log t\rfloor\}$. For each such n, consider the slice

$$E_{q,n} := \Big\{ K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \notin \mathcal{S}_{T_q},\ N_{a^\star(\mu)}(T_q) = n,\ n\,\mathrm{kl}_-\Big(\frac{S_{a^\star(\mu)}(T_q)}{n}, \mu^-_{a^\star(\mu)}\Big) > \frac{1+R_\mu}{2R_\mu}\log t \Big\}.$$

Then

$$\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} \cap \mathcal{M}_q^c = \bigcup_{n=1}^{\lfloor A\log t\rfloor} E_{q,n},$$

and the union is disjoint because each slice fixes the value of $N_{a^\star(\mu)}(T_q)$. If

$$\frac{1+R_\mu}{2R_\mu}\,\frac{\log t}{n} \le \mathrm{kl}(0, \mu^-_{a^\star(\mu)}),$$

let $x_n \in [0, \mu^-_{a^\star(\mu)})$ be the unique point with

$$\mathrm{kl}(x_n, \mu^-_{a^\star(\mu)}) = \frac{1+R_\mu}{2R_\mu}\,\frac{\log t}{n};$$

then on $E_{q,n}$ we have

$$\mathrm{kl}_-\Big(\frac{S_{a^\star(\mu)}(T_q)}{n}, \mu^-_{a^\star(\mu)}\Big) > \frac{1+R_\mu}{2R_\mu}\,\frac{\log t}{n} = \mathrm{kl}(x_n, \mu^-_{a^\star(\mu)}).$$

Since $\mathrm{kl}_-(x, \mu^-_{a^\star(\mu)}) = 0$ for $x > \mu^-_{a^\star(\mu)}$ and $x \mapsto \mathrm{kl}(x, \mu^-_{a^\star(\mu)})$ is strictly decreasing on $[0, \mu^-_{a^\star(\mu)}]$, this implies

$$\frac{S_{a^\star(\mu)}(T_q)}{n} < x_n \qquad \text{on } E_{q,n}.$$

Therefore

$$E_{q,n} \subseteq \Big\{N_{a^\star(\mu)}(T_q) = n,\ \frac{S_{a^\star(\mu)}(T_q)}{n} \le x_n\Big\}.$$

If instead

$$\frac{1+R_\mu}{2R_\mu}\,\frac{\log t}{n} > \mathrm{kl}(0, \mu^-_{a^\star(\mu)}),$$

then $E_{q,n} = \varnothing$, because $\mathrm{kl}_-(x, \mu^-_{a^\star(\mu)}) \le \mathrm{kl}(0, \mu^-_{a^\star(\mu)})$ for every $x \in [0,1]$. In the first case, let $(Y^\star_\ell)_{\ell \ge 1}$ denote the rewards obtained from the successive pulls of arm $a^\star(\mu)$; under $\mathbb{P}_\mu$, these are i.i.d. $\mathrm{Bern}(\mu_{a^\star(\mu)})$. On the event $\{N_{a^\star(\mu)}(T_q) = n\}$ we have

$$\frac{S_{a^\star(\mu)}(T_q)}{n} = \frac{1}{n}\sum_{\ell=1}^{n} Y^\star_\ell.$$

Therefore the inclusion above gives

$$E_{q,n} \subseteq \Big\{\frac{1}{n}\sum_{\ell=1}^{n} Y^\star_\ell \le x_n\Big\}.$$

Since $Y^\star_1 + \cdots + Y^\star_n \in \{0, \dots, n\}$ is integer-valued, this implies

$$E_{q,n} \subseteq \Big\{Y^\star_1 + \cdots + Y^\star_n \le m_n\Big\}, \qquad m_n := \lfloor n x_n \rfloor.$$

Lemma 2 therefore gives

$$\mathbb{P}_\mu(E_{q,n}) \le \exp\big(-n\,\mathrm{kl}(m_n/n, \mu_{a^\star(\mu)})\big).$$

Because

$$\frac{m_n}{n} \le x_n < \mu^-_{a^\star(\mu)} < \mu_{a^\star(\mu)},$$

and $x \mapsto \mathrm{kl}(x, \mu_{a^\star(\mu)})$ is strictly decreasing on $[0, \mu_{a^\star(\mu)})$, the definition of R_μ then gives

$$\mathbb{P}_\mu(E_{q,n}) \le \exp\big(-n\,\mathrm{kl}(m_n/n, \mu_{a^\star(\mu)})\big) \le \exp\big(-n\,\mathrm{kl}(x_n, \mu_{a^\star(\mu)})\big) \le t^{-(1+R_\mu)/2}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}_\mu\big(\{K \le T_q < \tau_{i,j+1},\ a^\star(\mu) \in \mathcal{U}\} \cap \mathcal{M}_q^c\big) &\le \sum_{n=1}^{\lfloor A \log t \rfloor} \mathbb{P}_\mu(E_{q,n}) \\
&\le \sum_{n=1}^{\lfloor A \log t \rfloor} t^{-(1+R_\mu)/2} \\
&\le (A \log t + 1)\, t^{-(1+R_\mu)/2}.
\end{aligned}$$

Since $R_\mu > 1$, the contribution from $\mathcal{M}_q$ bounded in Step (c)-2 is super-polynomially small, while the complementary contribution bounded in Step (c)-3 is polynomially small with exponent strictly

larger than 1. Therefore there exist constants $C_\Phi = C_\Phi(\mu, A) > 0$, $\gamma_\Phi = \gamma_\Phi(\mu, A) > 1$, and $t_\Phi = t_\Phi(\mu, A) \geq 3$ such that for all $t \geq t_\Phi$,

$$(c) \leq C_\Phi \, t^{-\gamma_\Phi}.$$

5. Conclusion. Collecting (12) and the bounds for (b) and (c), we obtain

$$\begin{aligned}\mathbb{P}_\mu(\widetilde{E}_{i,j,q}) &\leq (a) + (b) + (c) \\ &\leq 2(K-1)C_1 \, t^{-Ac_0} + \exp\Big(-\frac{t}{128K^2A^2(\log t)^{11}}\Big) \\ &\quad + \; C_\Phi \, t^{-\gamma_\Phi}.\end{aligned}$$

Set

$$\gamma := \frac{1}{2}\big(1 + \min\{Ac_0, \gamma_\Phi\}\big) > 1.$$

Then $\gamma < Ac_0$ and $\gamma < \gamma_\Phi$, so for all large enough t,

$$t^{-Ac_0} \leq t^{-\gamma}, \qquad t^{-\gamma_\Phi} \leq t^{-\gamma},$$

while the middle term $\exp\big(-t/[128K^2A^2(\log t)^{11}]\big)$ decays faster than any power of t. After enlarging thresholds and absorbing constants, we obtain

$$\mathbb{P}_\mu(\widetilde{E}_{i,j,q}) \leq C \, t^{-\gamma}$$

for all large enough t. This proves the lemma. □

F.1.3 Proof of Lemma 7

Proof of Lemma 7. Define

$$f(x) := \frac{\mathrm{kl}(x, \mu_{a^\star(\mu)})}{\mathrm{kl}(x, \mu^-_{a^\star(\mu)})} \qquad \text{for } x \in [0, \mu^-_{a^\star(\mu)}).$$

For every $x \in [0, \mu^-_{a^\star(\mu)})$, the map $q \mapsto \mathrm{kl}(x, q)$ is strictly increasing on $(x, 1)$ because

$$\partial_q \mathrm{kl}(x, q) = \frac{q - x}{q(1-q)} > 0 \qquad (q > x).$$

Since $\mu_{a^\star(\mu)} > \mu^-_{a^\star(\mu)} > x$, this yields

$$f(x) > 1 \qquad \text{for every } x \in [0, \mu^-_{a^\star(\mu)}).$$

Moreover,

$$\lim_{x \uparrow \mu^-_{a^\star(\mu)}} f(x) = +\infty,$$

because $\mathrm{kl}(x, \mu^-_{a^\star(\mu)}) \to 0$ while

$$\mathrm{kl}(x, \mu_{a^\star(\mu)}) \to \mathrm{kl}(\mu^-_{a^\star(\mu)}, \mu_{a^\star(\mu)}) > 0.$$

Therefore there exists $\varepsilon \in (0, \mu^-_{a^\star(\mu)})$ such that

$$f(x) \geq 2 \qquad \text{for all } x \in [\mu^-_{a^\star(\mu)} - \varepsilon, \mu^-_{a^\star(\mu)}).$$

On the compact interval $[0, \mu^-_{a^\star(\mu)} - \varepsilon]$, the function f is continuous, because $x \mapsto \mathrm{kl}(x, q)$ is continuous for every fixed $q \in (0, 1)$ and the denominator is strictly positive there. Hence f attains a minimum m on this interval. Since $f(x) > 1$ pointwise, we have $m > 1$. Consequently,

$$R_\mu = \inf_{x \in [0, \mu^-_{a^\star(\mu)})} f(x) \geq \min\{m, 2\} > 1.$$

□

F.1.4 Proof of Lemma 8

Proof of Lemma 8. We split according to whether the empirical mean is atypical, and then apply the conditional Beta-posterior tail estimate on the typical event. Fix an arm $k \in \{1, \dots, K\}$ and a stopping time τ with $K \le \tau < \infty$ almost surely.

$$\mathbb{P}_\mu\big(|\theta_{k,\tau+1} - \mu_k| > r_\mu,\ N_k(\tau) \ge A \log t\big) \le \mathbb{P}_\mu\big(|\hat{\mu}_k(\tau) - \mu_k| > r_\mu/2,\ N_k(\tau) \ge A \log t\big) + \mathbb{E}_\mu\left[\begin{array}{l}\mathbb{P}_\mu\big(|\theta_{k,\tau+1} - \mu_k| > r_\mu \mid \mathcal{F}_\tau\big) \\ \quad \mathbf{1}\{|\hat{\mu}_k(\tau) - \mu_k| \le r_\mu/2,\ N_k(\tau) \ge A \log t\}\end{array}\right].$$

Since $N_k(\tau) \ge A \log t$ implies $N_k(\tau) \ge \lceil A \log t \rceil$, Lemma 6 applied at the stopping time τ with radius $r_\mu/2$ gives

$$\mathbb{P}_\mu\big(|\hat{\mu}_k(\tau) - \mu_k| > r_\mu/2,\ N_k(\tau) \ge A \log t\big) \le C_{r_\mu/2}\, t^{-A r_\mu^2/2}.$$

On the event $\{|\hat{\mu}_k(\tau) - \mu_k| \le r_\mu/2\}$, Lemma 3 applied with $p = \mu_k$ and radius r_μ gives constants $\widetilde{C}_k, \widetilde{c}_k > 0$ such that

$$\mathbb{P}_\mu\big(|\theta_{k,\tau+1} - \mu_k| > r_\mu \mid \mathcal{F}_\tau\big) \le \widetilde{C}_k e^{-\widetilde{c}_k N_k(\tau)}.$$

Hence the second term is at most $\widetilde{C}_k t^{-A\widetilde{c}_k}$. Taking

$$c_0 := \min\Big\{\frac{r_\mu^2}{2}, \min_{1 \le k \le K} \widetilde{c}_k\Big\} > 0, \qquad C_1 := C_{r_\mu/2} + \max_{1 \le k \le K} \widetilde{C}_k,$$

and enlarging the lower threshold on t if needed, we obtain

$$\mathbb{P}_\mu\big(|\theta_{k,\tau+1} - \mu_k| > r_\mu,\ N_k(\tau) \ge A \log t\big) \le C_1\, t^{-A c_0},$$

uniformly over k and τ. This proves the lemma. □

F.1.5 Proof of Lemma 11

Proof of Lemma 11. We first prove the conditional-independence and fixed-posterior statements, and then derive the one-dimensional and pairwise anti-concentration bounds from the resulting Beta laws.

Fix n and work on the event $\{T_q = n\}$; then $s = T_q + r = n + r$ and $\mathcal{S}_{T_q} = \mathcal{S}_n$. Since

$$\Xi_{r-1} \cap \{T_q = n\} \subseteq \bigcap_{m=n+1}^{s-1} \{I_m \in \mathcal{S}_n\},$$

we have, for every $u \in \mathcal{U} = \{1, \dots, K\} \setminus \mathcal{S}_n$,

$$\begin{aligned} N_u(s-1) - N_u(n) &= \sum_{m=n+1}^{s-1} \mathbf{1}\{I_m = u\} = 0, \\ S_u(s-1) - S_u(n) &= \sum_{m=n+1}^{s-1} X_m\, \mathbf{1}\{I_m = u\} = 0 \end{aligned} \qquad \text{on } \Xi_{r-1} \cap \{T_q = n\}.$$

Hence

$$N_u(s-1) = N_u(n), \qquad S_u(s-1) = S_u(n) \qquad \text{on } \Xi_{r-1} \cap \{T_q = n\}.$$

Therefore the posterior parameters of arm u at time s are the same as at time n, and thus

$$\theta_{u,s} \mid \mathcal{F}_{s-1} \sim \mathrm{Beta}\big(\alpha_u + S_u(n), \beta_u + N_u(n) - S_u(n)\big) \qquad \text{on } \Xi_{r-1} \cap \{u \in \mathcal{U},\ T_q = n\}.$$

Moreover, on $\{T_q = n\}$,

$$\mathcal{H}_{q,r} = \sigma\Big(\mathcal{F}_{s-1}, \theta_{K+1}, \dots, \theta_{s-1}, \big(\theta_{k,s}\big)_{k \in \mathcal{S}_n}\Big) \subseteq \sigma\Big(\mathcal{F}_{s-1}, \theta_{K+1}, \dots, \theta_{s-1}, \big(U_{s,k}\big)_{k \in \mathcal{S}_n}\Big),$$

because each $\theta_{k,s}$ is $\sigma(\mathcal{F}_{s-1}, U_{s,k})$-measurable. By the deterministic-time enlarged-past independence recorded in Appendix C,

$$\big(U_{s,u}\big)_{u \in \mathcal{U}} \perp \sigma\Big(\mathcal{F}_{s-1}, \theta_{K+1}, \dots, \theta_{s-1}, \big(U_{s,k}\big)_{k \in \mathcal{S}_n}\Big) \qquad \text{on } \{T_q = n\},$$

and the family $\big(U_{s,u}\big)_{u\in\mathcal{U}}$ is mutually independent. Since each $\theta_{u,s}$ is a measurable function of $\mathcal{F}_{s-1}$ and $U_{s,u}$, it follows that on $\Xi_{r-1}\cap\{T_q=n\}$ the family $(\theta_{u,s})_{u\in\mathcal{U}}$ is conditionally independent given $\mathcal{H}_{q,r}$, and each arm u has the Beta law above.

If $a\in\mathcal{U}$, then on $\{T_q=n\}$,

$$\sigma\big(\mathcal{H}_{q,r},\theta_{a,s}\big)\subseteq\sigma\Big(\mathcal{F}_{s-1},\theta_{K+1},\dots,\theta_{s-1},\big(U_{s,k}\big)_{k\in\mathcal{S}_n\cup\{a\}}\Big).$$

Hence $\big(U_{s,u}\big)_{u\in\mathcal{U}\setminus\{a\}}$ remains independent of $\sigma(\mathcal{H}_{q,r},\theta_{a,s})$, so the remaining unsaturated arms are conditionally independent given $\mathcal{H}_{q,r},\theta_{a,s}$ and retain the same Beta laws. Since T_q is $\mathcal{H}_{q,r}$-measurable, summing over the disjoint slices $\{T_q=n\}$ yields both conditional-independence and fixed-posterior statements.

We now turn to anti-concentration. Recall that the density of $\mathrm{Beta}(a,b)$ with integers $a,b\geq 1$ is bounded by $a+b-1$: if $a=1$ or $b=1$ the density is monotone, while for $a,b\geq 2$ it can be written as $(a+b-1)$ times the binomial point mass $\mathbb{P}(\mathrm{Bin}(a+b-2,x)=a-1)$ and is therefore bounded by $a+b-1$. Consequently, for every arm u, on $\Xi_{r-1}\cap\{u\in\mathcal{U}\}$ the conditional law of $\theta_{u,s}$ given $\mathcal{H}_{q,r}$ has density bounded by

$$(\alpha_u+\beta_u-1)+N_u(T_q)\leq(\alpha_u+\beta_u-1)+A\log t,$$

because $u\in\mathcal{U}$ implies $N_u(T_q)<A\log t$. Since $\log t\geq 1$ for all large enough t, there exists a constant $C_0>0$, depending only on the prior, such that

$$(\alpha_u+\beta_u-1)+A\log t\leq C_0(1+A)\log t$$

for every arm u. An interval of radius δ_t has length $2\delta_t=2(\log t)^{-4}$, so

$$\begin{aligned}\mathbb{P}_\mu\big(|\theta_{u,s}-x|\leq\delta_t\mid\mathcal{H}_{q,r}\big)&\leq 2\delta_t\big((\alpha_u+\beta_u-1)+A\log t\big)\\&\leq\frac{C_0(1+A)}{(\log t)^3}\end{aligned}$$

for every $x\in(0,1)$, which proves (8). Exactly the same computation under conditioning on $\mathcal{H}_{q,r},\theta_{a,s}$ gives the conditioned one-dimensional bound from the statement.

Finally, let $u\neq v$ be two unsaturated arms. Using the tower property and then the conditioned one-dimensional bound with conditioning sigma-field $\sigma(\mathcal{H}_{q,r},\theta_{v,s})$, we obtain

$$\mathbb{P}_\mu\big(|\theta_{u,s}-\theta_{v,s}|\leq\delta_t\mid\mathcal{H}_{q,r}\big)=\mathbb{E}_\mu\Big[\mathbb{P}_\mu\big(|\theta_{u,s}-\theta_{v,s}|\leq\delta_t\mid\mathcal{H}_{q,r},\theta_{v,s}\big)\;\Big|\;\mathcal{H}_{q,r}\Big]\leq\frac{C_0(1+A)}{(\log t)^3}.$$

If a,u,v are distinct unsaturated arms, then conditioning further on $\theta_{a,s}$ gives

$$\begin{aligned}\mathbb{P}_\mu\big(|\theta_{u,s}-\theta_{v,s}|\leq\delta_t\mid\mathcal{H}_{q,r},\theta_{a,s}\big)&=\mathbb{E}_\mu\Big[\mathbb{P}_\mu\big(|\theta_{u,s}-\theta_{v,s}|\leq\delta_t\mid\mathcal{H}_{q,r},\theta_{a,s},\theta_{v,s}\big)\;\Big|\;\mathcal{H}_{q,r},\theta_{a,s}\Big]\\&\leq\frac{C_0(1+A)}{(\log t)^3},\end{aligned}$$

which is the conditioned pairwise bound used later in the proof of Lemma 10. □

F.1.6 Proof of Lemma 12

Proof of Lemma 12. Recall that

$$C^\star(x)=F(x,w^\star(x)),\qquad F(x,w)=\inf_{\lambda\in\mathrm{Alt}(x)}\sum_{k=1}^K w_k\,d(x_k,\lambda_k).$$

We first obtain an upper bound on $C^\star(x)$ in terms of each oracle weight $w_u^\star(x)$ with $u\in U$ by testing the oracle allocation against one-alternative instances.

Fix $u\in U$. If $u\neq a$, define $\lambda^{(u)}\in(0,1)^K$ by

$$\lambda_u^{(u)}=x_a,\qquad\lambda_k^{(u)}=x_k\quad\text{for } k\neq u.$$

Then $\lambda_u^{(u)} = \lambda_a^{(u)} = x_a$, so $\lambda^{(u)} \in \mathrm{Alt}(x)$ because $u \neq a$. Therefore

$$\begin{aligned} C^\star(x) = F(x, w^\star(x)) &\le \sum_{k=1}^{K} w_k^\star(x)\, d\big(x_k, \lambda_k^{(u)}\big) \\ &= w_u^\star(x)\, d(x_u, x_a). \end{aligned}$$

If instead $u = a$, define $\lambda^{(a)} \in (0,1)^K$ by

$$\lambda_a^{(a)} = x_b, \qquad \lambda_k^{(a)} = x_k \quad \text{for } k \neq a.$$

Then $\lambda_b^{(a)} = \lambda_a^{(a)} = x_b$, so $\lambda^{(a)} \in \mathrm{Alt}(x)$ because $b \neq a$. Hence

$$\begin{aligned} C^\star(x) = F(x, w^\star(x)) &\le \sum_{k=1}^{K} w_k^\star(x)\, d\big(x_k, \lambda_k^{(a)}\big) \\ &= w_a^\star(x)\, d(x_a, x_b). \end{aligned}$$

Thus, for every $u \in U$,

$$C^\star(x) \le w_u^\star(x)\, \kappa_u,$$

where $\kappa_u = d(x_u, x_a)$ if $u \neq a$, and $\kappa_a = d(x_a, x_b)$.

We next bound each κ_u from above. In both cases, the second argument of the KL divergence is either x_a or x_b, hence lies in $[1/t, 1 - 1/t]$ by assumption. Therefore, for every $p \in (0,1)$ and every $q \in [1/t, 1 - 1/t]$,

$$\begin{aligned} d(p,q) &= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \\ &\le -p \log q - (1-p) \log(1-q) \\ &\le \max\{-\log q,\, -\log(1-q)\} \\ &\le \log t. \end{aligned}$$

Hence $\kappa_u \le \log t$ for every $u \in U$, and so

$$C^\star(x) \le w_u^\star(x)\ \log t \qquad \text{for every } u \in U.$$

We now derive a lower bound on $C^\star(x)$. Let $\bar{w} := (1/K, \dots, 1/K)$. We claim that

$$F(x, \bar{w}) \ge \frac{\delta_t^2}{2K}.$$

To prove this, fix any $\lambda \in \mathrm{Alt}(x)$. By definition of $\mathrm{Alt}(x)$, there exists some $j \neq a$ such that $\lambda_j \ge \lambda_a$. Since b is a second-best arm of x, we have $x_j \le x_b$, and therefore

$$x_a - x_j \ge x_a - x_b \ge \delta_t.$$

Suppose for contradiction that both

$$|\lambda_a - x_a| < \frac{\delta_t}{2} \qquad \text{and} \qquad |\lambda_j - x_j| < \frac{\delta_t}{2}.$$

Then

$$\lambda_a > x_a - \frac{\delta_t}{2} \ge x_j + \frac{\delta_t}{2} > \lambda_j,$$

which contradicts $\lambda_j \ge \lambda_a$. Thus at least one of the two inequalities

$$|\lambda_a - x_a| \ge \frac{\delta_t}{2}, \qquad |\lambda_j - x_j| \ge \frac{\delta_t}{2}$$

must hold. By Pinsker's inequality,

$$d(p,q) \ge 2(p-q)^2 \qquad \text{for all } p, q \in (0,1),$$

so for at least one of the indices a or j we obtain

$$d(x_k, \lambda_k) \ge 2\Big(\frac{\delta_t}{2}\Big)^2 = \frac{\delta_t^2}{2}.$$

Since all terms are nonnegative,

$$\sum_{k=1}^{K} \bar{w}_k \, d(x_k, \lambda_k) \geq \frac{1}{K} \cdot \frac{\delta_t^2}{2} = \frac{\delta_t^2}{2K}.$$

Because this bound is uniform over $\lambda \in \mathrm{Alt}(x)$, taking the infimum over λ yields

$$F(x, \bar{w}) \geq \frac{\delta_t^2}{2K}.$$

Consequently,

$$C^\star(x) \geq F(x, \bar{w}) \geq \frac{\delta_t^2}{2K}.$$

Combining the upper and lower bounds, we find that for every $u \in U$,

$$w_u^\star(x) \geq \frac{C^\star(x)}{\log t} \geq \frac{\delta_t^2}{2K \log t} = \frac{1}{2K(\log t)^9}.$$

Summing this inequality over $u \in U$ gives (9). □

F.1.7 Proof of Lemma 13

Proof of Lemma 13. For $n = 0$, the claim is immediate because $\mathrm{Beta}(\alpha, \beta)$ has a density strictly positive on $(0, 1)$. After shrinking c_τ if needed, we may therefore assume $n \geq 1$.

By the same Beta–Binomial identity used earlier,

$$\mathbb{P}(\theta \geq \tau) = \mathbb{P}\big(\mathrm{Bin}(n + \alpha + \beta - 1, \tau) \leq \alpha + s - 1\big).$$

Write this binomial variable as the sum of an n-trial binomial and an independent $(\alpha + \beta - 1)$-trial binomial. Since the latter is at most $\alpha - 1$ with positive probability depending only on τ, α, β, the prior contributes only a fixed multiplicative loss. Hence it is enough, up to decreasing c_τ, to lower bound

$$\mathbb{P}\big(\mathrm{Bin}(n, \tau) \leq s\big).$$

By the exact formulas at $m = 0$ and $m = n$, and by Stirling's bounds for $1 \leq m \leq n - 1$, after decreasing c_τ if necessary we have for every $m \in \{0, \dots, n\}$,

$$\mathbb{P}\big(\mathrm{Bin}(n, \tau) = m\big) \geq c_\tau \, (n + 1)^{-1/2} \exp\Big(-n \, \mathrm{kl}\big(m/n, \tau\big)\Big). \tag{30}$$

Indeed, for $1 \leq m \leq n - 1$, Stirling gives

$$\binom{n}{m} \geq c_\tau \, (n + 1)^{-1/2} \Big(\frac{n}{m}\Big)^m \Big(\frac{n}{n - m}\Big)^{n-m},$$

and therefore

$$\begin{aligned}
\mathbb{P}\big(\mathrm{Bin}(n, \tau) = m\big) &= \binom{n}{m} \tau^m (1 - \tau)^{n-m} \\
&\geq c_\tau \, (n + 1)^{-1/2} \Big(\frac{n\tau}{m}\Big)^m \Big(\frac{n(1 - \tau)}{n - m}\Big)^{n-m} \\
&= c_\tau \, (n + 1)^{-1/2} \exp\Big(-n \, \mathrm{kl}\big(m/n, \tau\big)\Big).
\end{aligned}$$

If $s/n \leq \tau$, take $m = s$ in (30). If $s/n > \tau$, take $m_n := \lfloor n\tau \rfloor \leq s$; then $m_n/n \in [\tau - 1/n, \tau]$, so m_n is a representative point near the threshold where the lower tail still contributes at least one local mass term. Since $n \, \mathrm{kl}(m_n/n, \tau)$ stays uniformly bounded, another decrease of c_τ gives

$$\mathbb{P}\big(\mathrm{Bin}(n, \tau) \leq s\big) \geq c_\tau \, (n + 1)^{-1/2}.$$

In both cases this is at least

$$c_\tau \, (n + 1)^{-1/2} \exp\Big(-n \, \mathrm{kl}_-\big(s/n, \tau\big)\Big),$$

and combining this with the first paragraph proves the lemma. □

F.1.8 Proof of Lemma 14

Proof of Lemma 14. On $\mathcal{M}_q$, Lemma 13 applied with threshold $\mu^-_{a^\star(\mu)}$ gives, for a constant $c > 0$ depending only on μ,

$$\Phi_q \geq c\big(N_{a^\star(\mu)}(T_q)+1\big)^{-1/2} \exp\!\left(-N_{a^\star(\mu)}(T_q)\,\mathrm{kl}_-\Big(\frac{S_{a^\star(\mu)}(T_q)}{N_{a^\star(\mu)}(T_q)}, \mu^-_{a^\star(\mu)}\Big)\right).$$

Since $a^\star(\mu) \notin \mathcal{S}_{T_q}$ implies $N_{a^\star(\mu)}(T_q) \leq \lfloor A\log t \rfloor$, on $\mathcal{M}_q$ we obtain

$$\Phi_q \geq c\,(A\log t+1)^{-1/2} t^{-(1+R_\mu)/(2R_\mu)}.$$

The claim follows after renaming the constant. □

G Proof of Lemma 5

We combine Proposition 1 with the conditional posterior tail bound from Lemma 3 to turn sample-size control into time-polynomial control.

Proof. Fix $r_0 \in (0, r_\mu]$. For each arm i, let $(C_{0,i}, c_{0,i})$ be the constants from Lemma 3 applied with $p = \mu_i$ and radius r_0 (and prior parameters (α_i, β_i)). Define $c_p := \min_{1\leq i\leq K} c_{0,i} > 0$ and set

$$c_{\min} := \min\Big\{\frac{r_0^2}{2},\ c_p\Big\} > 0.$$

Choose $A \geq 2/c_{\min}$ large enough that Proposition 1 applies with parameter $A+1$, so in particular $c_{\min}A \geq 2 > 1$. The extra $+1$ is only technical slack: it lets us compare the threshold $A\log t$ with a logarithmic threshold evaluated at time $t-1$. For $t \geq 3$ define $\mathcal{N}_t := \{\min_i N_i(t-1) \geq A\log t\}$. Then by a union bound over i,

$$\mathbb{P}_\mu\big(\|\theta_t - \mu\|_\infty > r_0\big) \leq \mathbb{P}_\mu(\mathcal{N}_t^c) + \sum_{i=1}^K \mathbb{P}_\mu\big(|\theta_{i,t}-\mu_i| > r_0,\ \mathcal{N}_t\big).$$

Step 1: bound $\mathbb{P}_\mu(\mathcal{N}_t^c)$. Since $\log t = \log(t-1) + \log(1+1/(t-1)) \leq \log(t-1) + 1/(t-1)$, there exists $t_1 = t_1(A) \geq 3$ such that for all $t \geq t_1$,

$$A\log t \leq (A+1)\log(t-1).$$

Hence, for all $t \geq t_1$,

$$\mathcal{N}_t^c = \Big\{\min_i N_i(t-1) < A\log t\Big\} \subseteq \Big\{\min_i N_i(t-1) < (A+1)\log(t-1)\Big\}.$$

Applying Proposition 1 at time $t-1$ with parameter $A+1$, we obtain constants $C_n = C_n(\mu, A+1) > 0$ and $q_n = q_n(\mu, A+1) > 1$ such that for all t large enough,

$$\mathbb{P}_\mu(\mathcal{N}_t^c) \leq C_n\, t^{-q_n}.$$

Step 2: bound the posterior-draw terms on $\mathcal{N}_t$. Fix an arm i and let $n_t := \lceil A\log t\rceil$. On $\mathcal{N}_t$, the sample size of every arm is already logarithmic in t, so the exponential sample-size estimates from the one-arm lemmas become polynomial in t. We decompose

$$\begin{aligned}\mathbb{P}_\mu\big(|\theta_{i,t}-\mu_i| > r_0,\ \mathcal{N}_t\big) &\leq \mathbb{P}_\mu\big(|\hat\mu_i(t-1)-\mu_i| > r_0/2,\ N_i(t-1) \geq n_t\big)\\ &\quad + \mathbb{E}_\mu\big[\mathbb{P}_\mu\big(|\theta_{i,t}-\mu_i| > r_0 \mid \mathcal{F}_{t-1}\big)\,\mathbf{1}\{|\hat\mu_i(t-1)-\mu_i| \leq r_0/2,\ \mathcal{N}_t\}\big].\end{aligned}$$

First term. Applying Lemma 6 with $\tau = t-1$, $u = r_0/2$, and $n = n_t$ yields

$$\mathbb{P}_\mu\big(|\hat\mu_i(t-1)-\mu_i| > r_0/2,\ N_i(t-1) \geq n_t\big) \leq C_{r_0/2}\,\exp\!\Big(-\frac{r_0^2}{2}n_t\Big) \leq C_{r_0/2}\,t^{-(r_0^2/2)A},$$

where $C_{r_0/2}$ is the constant from Lemma 6.

Second term. On $\{|\hat{\mu}_i(t-1) - \mu_i| \le r_0/2\}$ we have $|S_i(t-1)/N_i(t-1) - \mu_i| \le r_0/2$ and

$$\theta_{i,t} \mid \mathcal{F}_{t-1} \sim \mathrm{Beta}\big(\alpha_i + S_i(t-1),\ \beta_i + N_i(t-1) - S_i(t-1)\big).$$

Therefore, by Lemma 3, on $\{|\hat{\mu}_i(t-1) - \mu_i| \le r_0/2\}$ we have

$$\mathbb{P}_\mu\big(|\theta_{i,t} - \mu_i| > r_0 \mid \mathcal{F}_{t-1}\big) \le C_{0,i}\, e^{-c_{0,i} N_i(t-1)}.$$

On $\mathcal{N}_t$, $N_i(t-1) \ge A \log t$, hence the second term is bounded by $C_{0,i} t^{-c_{0,i} A}$.

Conclusion of Step 2. Combining the bounds above gives a constant $C_i' > 0$ such that for all large enough t,

$$\mathbb{P}_\mu\big(|\theta_{i,t} - \mu_i| > r_0,\ \mathcal{N}_t\big) \le C_i'\, t^{-c_{\min} A}.$$

Summing over i yields $\sum_i \mathbb{P}_\mu(|\theta_{i,t} - \mu_i| > r_0,\ \mathcal{N}_t) \le C'\, t^{-c_{\min} A}$ for a constant $C' > 0$.

Conclusion. Let $p := \min\{q_n,\ c_{\min} A\} > 1$ and absorb constants to get the claim. □

H Numerical experiments

H.1 Experimental setup and stopping thresholds

We compare Posterior Tracking with D-Tracking from Track-and-Stop (Garivier and Kaufmann, 2016), TTTS with $\beta = 1/2$ and TTTS with oracle β (Russo, 2020; Shang et al., 2020; Jourdan et al., 2022), and BC-TE (Lee et al., 2024). All methods use the Bernoulli GLR statistic and the same five values of δ: $\delta \in \{10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}\}$. The oracle-$\beta$ TTTS baseline uses the best-arm coordinate of $w^\star(\mu)$ and is included only as an oracle benchmark.

The main comparison uses the empirical threshold

$$\beta_{\mathrm{emp}}(t, \delta) := \log\left(\frac{\log(\max\{t, 2\}) + 1}{\delta}\right), \tag{31}$$

as in numerical comparisons following Garivier and Kaufmann (2016); similar thresholds containing a log–log term are also used by Wang et al. (2021) and Tabata et al. (2023). This threshold is less conservative than the threshold $\beta(t, \delta)$ in (2), but (31) itself has no finite-δ correctness guarantee.

For each instance–method pair, we generated 100 independent trajectories. Because the sampling paths do not depend on δ, we recorded the stopping time under β_{emp} for each of the five values of δ along every trajectory. We also extended the prespecified subset consisting of the first 30 trajectories for each instance–method pair until the stopping condition using $\beta(t, \delta)$ was met for all five values of δ. Thus the two thresholds were evaluated on identical trajectories. The maximum horizon was 2×10^6 samples. Every stopping time using β_{emp}, and every stopping time using $\beta(t, \delta)$ in the 30-trajectory subset, was at most this horizon.

Table 2 defines the three main instances.

Table 2: Experimental instances.

Label	Arm means	Purpose
OSC	$(0.60, 0.55, 0.35, 0.25, 0.15)$	One small gap and three larger gaps
K6-low	$(0.60, 0.582, 0.20, 0.20, 0.20, 0.20)$	One small gap and four larger gaps
K10-eq	$(0.50, 0.30, \ldots, 0.30)$ (10 arms)	Nine suboptimal arms with the same mean

For each reported group, the standard error is the sample standard deviation divided by the square root of the number of trajectories. The random seed was 20260426. The simulations for the main results and the comparison of stopping thresholds used 12 CPU workers and took 13,876.94 seconds (approximately 3 hours 51 minutes). The experiments were run on a MacBook Pro with an Apple M4 Max chip and 64 GB of memory, using Python 3.9.6, NumPy 2.0.2, SciPy 1.13.1, and macOS 26.5.2; no GPU was used.

H.2 Oracle allocation computation

The experiments use the one-dimensional characterization of the Bernoulli BAI oracle allocation from Garivier and Kaufmann (2016), rather than a generic K-dimensional constrained optimizer. Given an input θ with unique best arm a, write $p = \theta_a$, $q_j = \theta_j$, and

$$I_j(r) := \mathrm{kl}\left(p, \frac{p + rq_j}{1+r}\right) + r\,\mathrm{kl}\left(q_j, \frac{p + rq_j}{1+r}\right), \qquad j \neq a.$$

For each candidate information level h, the code solves the $K-1$ monotone equations $I_j(r_j(h)) = h$ using a vectorized safeguarded Newton method, with a bisection step whenever a Newton step falls outside the current bracket. It then solves

$$1 + \sum_{j \neq a} r_j(h) - h \sum_{j \neq a} \frac{1}{I_j'(r_j(h))} = 0, \qquad I_j'(r) = \mathrm{kl}\left(q_j, \frac{p + rq_j}{1+r}\right),$$

and reconstructs

$$w_a^\star = \frac{1}{1 + \sum_{j \neq a} r_j(h^\star)}, \qquad w_j^\star = r_j(h^\star) w_a^\star.$$

Each evaluation of the left-hand side of this scalar equation costs $O(K)$ given the fixed number of vectorized inner iterations used here. Since the maximum number of outer iterations is also independent of K, the arithmetic cost is linear in K at these fixed numerical settings.

To validate the implementation numerically, we compared the allocations returned by this solver and by the generic SLSQP solver on the five instances specified in this section (including the $K = 20$ and $K = 40$ instances below) and on 30 pseudorandomly generated mean vectors, each with a unique best arm, using seed 20260805 ($K \in \{3, 5, 10\}$ and best-arm gap greater than 0.03). Across these 35 inputs, the maximum ℓ_∞-distance between the allocation vectors returned by the two solvers was 2.23×10^{-6}, and the maximum relative difference between the values of $C^\star(\theta)$ returned by the two solvers was 3.26×10^{-8}.

The theoretical analysis uses the exact allocation $w^\star(\theta_t)$ at every round. The numerical comparisons above provide a check on the tested inputs but do not establish that a fixed numerical tolerance preserves the exploration and asymptotic expected-sample-complexity guarantees. Since oracle coordinates can be arbitrarily small, a fixed additive error bound alone does not ensure the selection probabilities required by the exploration proof. Extending the analysis to approximate allocations would require controlling both the probabilities of sampling under-sampled arms and the effect of approximation errors on allocation stabilization. We leave this analysis to future work.

H.3 Main results from 100 trajectories

Tables 3, 4, and 5 report the β_{emp} results from 100 trajectories for every instance–method pair. The Samples column reports the empirical mean of τ_δ with standard errors; allocation error is $\|N(\tau_\delta)/\tau_\delta - w^\star(\mu)\|_\infty$; best-arm fraction is $N_{a^\star(\mu)}(\tau_\delta)/\tau_\delta$; and correctness is the empirical fraction of runs that recommend the true best arm. The remaining columns report empirical means. Within each value of δ, the smallest mean stopping time in the Samples column is shown in bold.

Table 3: Results for OSC (one suboptimal arm with mean close to that of the best arm). The smallest mean stopping time in the Samples column for each value of δ is shown in bold.

δ	Method	Samples	Alloc. error	Best-arm frac.	Correctness
10^{-6}	Posterior Tracking	12558 ± 448	0.0123	0.487	1.00
	D-Tracking	13337 ± 486	0.0193	0.474	1.00
	TTTS $\beta = 1/2$	13306 ± 431	0.0094	0.497	1.00
	TTTS oracle β	**12161 ± 443**	0.0081	0.491	1.00
	BC-TE	13017 ± 500	0.0048	0.491	1.00
10^{-7}	Posterior Tracking	14281 ± 480	0.0109	0.487	1.00
	D-Tracking	15525 ± 513	0.0160	0.477	1.00
	TTTS $\beta = 1/2$	15107 ± 466	0.0091	0.497	1.00
	TTTS oracle β	**13647 ± 460**	0.0076	0.491	1.00
	BC-TE	15028 ± 533	0.0047	0.491	1.00
10^{-8}	Posterior Tracking	16364 ± 492	0.0099	0.488	1.00
	D-Tracking	17591 ± 533	0.0141	0.480	1.00
	TTTS $\beta = 1/2$	16899 ± 485	0.0085	0.497	1.00
	TTTS oracle β	**15466 ± 506**	0.0074	0.491	1.00
	BC-TE	16781 ± 569	0.0043	0.492	1.00
10^{-9}	Posterior Tracking	18497 ± 553	0.0091	0.488	1.00
	D-Tracking	19701 ± 586	0.0125	0.481	1.00
	TTTS $\beta = 1/2$	18822 ± 507	0.0082	0.498	1.00
	TTTS oracle β	**17594 ± 547**	0.0072	0.491	1.00
	BC-TE	18775 ± 621	0.0042	0.492	1.00
10^{-10}	Posterior Tracking	20191 ± 571	0.0086	0.489	1.00
	D-Tracking	21744 ± 630	0.0112	0.483	1.00
	TTTS $\beta = 1/2$	20874 ± 540	0.0081	0.498	1.00
	TTTS oracle β	**19429 ± 560**	0.0067	0.491	1.00
	BC-TE	20492 ± 632	0.0038	0.492	1.00

Table 4: Results for K6-low (one suboptimal arm with mean close to that of the best arm and four arms with lower means). The smallest mean stopping time in the Samples column for each value of δ is shown in bold.

δ	Method	Samples	Alloc. error	Best-arm frac.	Correctness
10^{-6}	Posterior Tracking	98403 ± 3173	0.0023	0.499	1.00
	D-Tracking	107406 ± 4002	0.0349	0.465	1.00
	TTTS $\beta = 1/2$	97444 ± 3758	0.0020	0.500	1.00
	TTTS oracle β	**86429 ± 3097**	0.0017	0.499	1.00
	BC-TE	93557 ± 3806	0.0006	0.499	1.00
10^{-7}	Posterior Tracking	114143 ± 3276	0.0019	0.499	1.00
	D-Tracking	120881 ± 4234	0.0317	0.468	1.00
	TTTS $\beta = 1/2$	112244 ± 4006	0.0018	0.500	1.00
	TTTS oracle β	**98776 ± 3289**	0.0016	0.499	1.00
	BC-TE	107036 ± 4120	0.0005	0.499	1.00
10^{-8}	Posterior Tracking	128135 ± 3425	0.0018	0.499	1.00
	D-Tracking	133784 ± 4174	0.0288	0.471	1.00
	TTTS $\beta = 1/2$	126009 ± 4216	0.0016	0.500	1.00
	TTTS oracle β	**111248 ± 3340**	0.0015	0.499	1.00
	BC-TE	121396 ± 4254	0.0005	0.499	1.00
10^{-9}	Posterior Tracking	142954 ± 3553	0.0017	0.499	1.00
	D-Tracking	146166 ± 4705	0.0265	0.473	1.00
	TTTS $\beta = 1/2$	143083 ± 4574	0.0015	0.500	1.00
	TTTS oracle β	**123818 ± 3394**	0.0013	0.499	1.00
	BC-TE	134971 ± 4476	0.0004	0.499	1.00
10^{-10}	Posterior Tracking	157317 ± 3772	0.0016	0.499	1.00
	D-Tracking	160337 ± 4882	0.0243	0.475	1.00
	TTTS $\beta = 1/2$	157962 ± 4531	0.0014	0.500	1.00
	TTTS oracle β	**136942 ± 3477**	0.0012	0.499	1.00
	BC-TE	151732 ± 4558	0.0004	0.499	1.00

Table 5: Results for K10-eq (ten arms with equal suboptimal means). The smallest mean stopping time in the Samples column for each value of δ is shown in bold.

δ	Method	Samples	Alloc. error	Best-arm frac.	Correctness
10^{-6}	Posterior Tracking	4075 ± 84	0.0583	0.297	1.00
	D-Tracking	3114 ± 63	0.0597	0.258	1.00
	TTTS $\beta = 1/2$	3677 ± 60	0.2298	0.476	1.00
	TTTS oracle β	**2971 ± 56**	0.0539	0.255	1.00
	BC-TE	3166 ± 77	0.0995	0.163	1.00
10^{-7}	Posterior Tracking	4658 ± 87	0.0543	0.293	1.00
	D-Tracking	3563 ± 66	0.0557	0.257	1.00
	TTTS $\beta = 1/2$	4295 ± 64	0.2337	0.480	1.00
	TTTS oracle β	**3431 ± 65**	0.0514	0.254	1.00
	BC-TE	3710 ± 86	0.0977	0.158	1.00
10^{-8}	Posterior Tracking	5152 ± 91	0.0514	0.290	1.00
	D-Tracking	3993 ± 69	0.0530	0.256	1.00
	TTTS $\beta = 1/2$	4836 ± 67	0.2359	0.482	1.00
	TTTS oracle β	**3933 ± 68**	0.0489	0.253	1.00
	BC-TE	4272 ± 96	0.0989	0.155	1.00
10^{-9}	Posterior Tracking	5593 ± 91	0.0494	0.288	1.00
	D-Tracking	4447 ± 71	0.0496	0.255	1.00
	TTTS $\beta = 1/2$	5378 ± 73	0.2377	0.484	1.00
	TTTS oracle β	**4392 ± 73**	0.0455	0.252	1.00
	BC-TE	4827 ± 100	0.1005	0.151	1.00
10^{-10}	Posterior Tracking	6192 ± 96	0.0467	0.286	1.00
	D-Tracking	4922 ± 76	0.0461	0.254	1.00
	TTTS $\beta = 1/2$	5994 ± 75	0.2392	0.485	1.00
	TTTS oracle β	**4890 ± 80**	0.0438	0.252	1.00
	BC-TE	5347 ± 110	0.1020	0.149	1.00

Across these 15 combinations of instance and δ, oracle-β TTTS has the smallest observed mean stopping time. All 100 trajectories in every table entry stopped and recommended the true best arm. The reported correctness 1.00 therefore means only that no error was observed in this finite simulation sample; it does not empirically certify error probabilities as small as 10^{-6}–10^{-10}.

H.4 Comparison of stopping thresholds

For each instance–method pair, we evaluated both thresholds, β_{emp} and $\beta(t, \delta)$ from (2), on the same prespecified 30 trajectories. Table 6 reports

$$\frac{\text{mean stopping time using } \beta(t, \delta)}{\text{mean stopping time using } \beta_{\mathrm{emp}}},$$

not the mean of the ratios computed separately for the 30 trajectories.

Table 6: Ratio of mean stopping times using $\beta(t, \delta)$ and β_{emp}, evaluated on the same trajectories.

Instance	Method	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
OSC	Posterior Tracking	3.43	3.20	2.88	2.62	2.51
	D-Tracking	2.97	2.81	2.60	2.40	2.29
	TTTS $\beta = 1/2$	3.18	2.94	2.73	2.56	2.38
	TTTS oracle β	3.55	3.28	3.02	2.76	2.60
	BC-TE	3.26	3.11	3.01	2.81	2.62

K6-low	Posterior Tracking	3.64	3.19	2.94	2.73	2.50
	D-Tracking	2.88	2.70	2.60	2.50	2.37
	TTTS $\beta = 1/2$	3.18	2.94	2.75	2.56	2.37
	TTTS oracle β	3.80	3.53	3.30	3.16	2.96
	BC-TE	3.32	3.09	2.91	2.75	2.50
K10-eq	Posterior Tracking	2.74	2.55	2.44	2.39	2.24
	D-Tracking	3.23	2.97	2.83	2.67	2.50
	TTTS $\beta = 1/2$	3.40	3.06	2.88	2.71	2.52
	TTTS oracle β	3.37	3.09	2.79	2.60	2.48
	BC-TE	3.57	3.22	2.94	2.72	2.63

This ratio ranges from 2.24 to 3.80 over all methods and settings and from 2.24 to 3.64 for Posterior Tracking. For Posterior Tracking it decreases as δ decreases on each of the three instances. For 6 of the 15 combinations of instance and δ, the method attaining the smallest mean stopping time differs between the two thresholds: OSC at 10^{-6} and K6-low at all five values of δ. No error was observed among the 30 continued trajectories in any setting, but this again does not replace the theorem's δ-correctness guarantee. These results quantify sensitivity to the choice of stopping threshold rather than establishing a method ranking that is invariant to its choice.

H.5 Comparison for different numbers of arms

To vary the number of arms having small oracle weights while keeping the means of the best and second-best arms fixed, we considered

$$\mu^{(K)} = (0.60, 0.582, 0.20, \ldots, 0.20), \qquad K \in \{6, 20, 40\}.$$

The corresponding characteristic times are $T^\star = 5979.9$, 6021.8, and 6081.6, respectively, a range of less than 2%. For comparison, we also included K10-eq, $(0.50, 0.30, \ldots, 0.30)$, an instance with $K = 10$ arms, equal suboptimal means, and $T^\star = 192.8$. For every instance we ran 100 paired Posterior Tracking and D-Tracking trajectories, using the same reward stream for each arm within each pair and the same threshold β_{emp}. Figure 1 shows the results for $\mu^{(K)}$, and Table 7 reports the results for K10-eq, at $\delta = 10^{-6}$ and 10^{-10}; stopping times for all five values of δ were recorded on every trajectory.

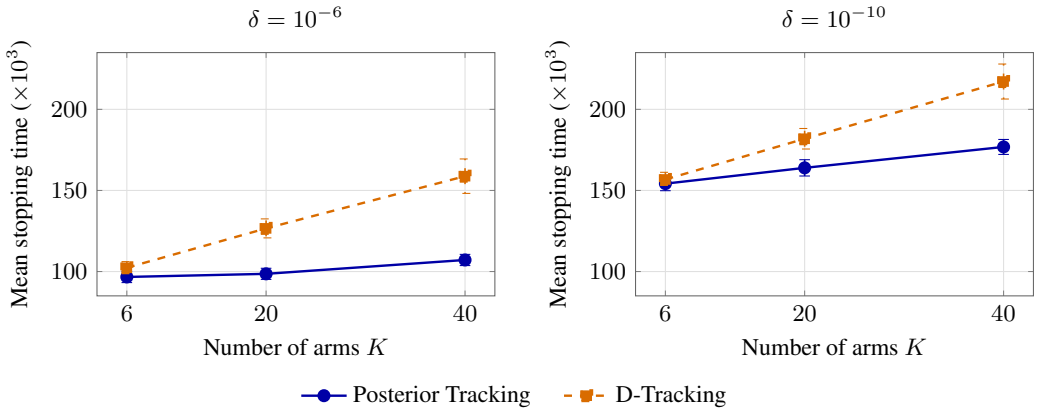

Figure 1: Mean stopping times on $\mu^{(K)}$ using β_{emp}, over 100 paired trajectories. Error bars show one standard error for each method. Both panels use the same vertical scale.

Table 7: K10-eq: mean stopping time $\pm$ standard error over 100 paired trajectories.

δ	Posterior Tracking	D-Tracking
10^{-6}	4095 ± 76	3016 ± 65
10^{-10}	6245 ± 115	4794 ± 84

Within the family $\{\mu^{(K)} : K \in \{6, 20, 40\}\}$, the observed difference in mean stopping time (D-Tracking minus Posterior Tracking) increases with K at both displayed values of δ, whereas D-Tracking has the smaller mean stopping time on K10-eq. Because the sampling rules also differ in other respects, this comparison does not isolate the effect of D-Tracking's explicit forced-exploration rule and does not establish uniform dominance or a lower-order improvement. These paired simulations took 22,230.54 seconds (approximately 6 hours 11 minutes) on the compute environment above, using 12 CPU workers and a maximum horizon of 10^6.

H.6 Computation time for the oracle allocation

We benchmarked the one-dimensional oracle-allocation solver described above using a single process on the Apple M4 Max machine. For each $K \in \{5, 6, 10, 20, 40, 100, 200\}$, the benchmark used 1,800 mean vectors, each with a unique best arm, and five repetitions, for 9,000 timed calls per K. Table 8 reports the median and 95th percentile.

Table 8: Computation time for the oracle allocation, in milliseconds.

K	Median	95th percentile
5	2.71	3.24
6	2.74	3.28
10	2.80	3.44
20	2.88	3.73
40	2.99	3.91
100	3.37	4.94
200	4.08	5.59

A linear regression of median running time on K yields an R^2 of 0.998 and a slope of 6.83 μs per arm over the tested range. The benchmark took 204.65 seconds (approximately 3 minutes 25 seconds). These measurements exclude input generation, posterior sampling, the categorical arm draw, reward generation, state updates, GLR computation, stopping checks, validation, file I/O, and multiprocessing overhead; they are not end-to-end algorithm runtimes.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the main scope and contributions of the paper: posterior tracking for frequentist fixed-confidence Bernoulli best-arm identification, Bernoulli GLR stopping, δ-correctness, and asymptotic optimality. These claims are formalized in Theorems 1, 2, and 3, together with the scope restrictions summarized in the "Limitations" section.

   Guidelines:

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

2. **Limitations**

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

Answer: [Yes]

Justification: The paper includes a dedicated "Limitations" section in the main text, which states that the theory is proved only for frequentist fixed-confidence K-armed Bernoulli BAI, that the expected-sample-complexity optimality guarantee is asymptotic as $\delta \to 0$, and that the finite-δ sample-complexity constants are not made explicit.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The model assumptions are stated in Section 2, and the main results are stated in Section 4. Proofs are provided in Sections 5 and 6 and in the appendix, including Appendix C.5.1 and Section E.

   Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The experiments are auxiliary to the theoretical contributions. Appendix H specifies the Bernoulli instances, methods compared, five values of δ, the generation of 100 independent trajectories for each instance–method pair, the evaluation of both stopping thresholds on the same prespecified subset of 30 trajectories, the paired finite-time comparison for the family $\{\mu^{(K)}\}$, and the oracle-allocation benchmark. It also specifies the reported means and standard errors and the computing environment. The supplemental material includes the main experiment script and run instructions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: The supplemental material includes the simulation code together with run instructions. The experiments use synthetic Bernoulli instances and do not rely on external datasets.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Appendix H describes the Bernoulli instances, methods compared, the five values of δ, 100 independent trajectories for each instance–method pair, the recording of the stopping time using β_{emp} for each value of δ along every trajectory, evaluation of the threshold $\beta(t, \delta)$ on the same prespecified subset of 30 trajectories, coupling of the Posterior Tracking and D-Tracking trajectories in each pair through a common reward sequence for each arm in the finite-time comparison, and all reported metrics. There are no training or test splits because the experiments are simulations on synthetic bandit instances.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The main experiment tables report mean stopping times with standard errors over 100 independent trajectories for each instance–method pair; each standard error is the sample standard deviation divided by the square root of the number of trajectories. The two stopping thresholds are evaluated on the same prespecified 30 trajectories, and the finite-time comparison uses 100 paired trajectories.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix H reports that the simulations were run on a MacBook Pro with an Apple M4 Max chip and 64 GB of memory, running macOS 26.5.2, using parallel CPU workers with no GPU acceleration. The simulations for the main results and the comparison of stopping thresholds took approximately 3 hours 51 minutes, the paired finite-time comparison took approximately 6 hours 11 minutes, and the single-process oracle-allocation benchmark took approximately 3 minutes 25 seconds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This is a theoretical study of pure-exploration bandit algorithms and does not involve human subjects, private data, or high-risk released assets. We are not aware of any aspect of the research that conflicts with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The conclusion contains a dedicated "Broader impacts" paragraph that discusses both possible benefits, such as reducing unnecessary exposure to inferior actions in online experimentation, and possible harms, such as reinforcing exposure biases or feedback loops in recommendation and advertising settings.

Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release high-risk models, datasets, or other assets that would require special safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper does not rely on external code, datasets, or pretrained models as research assets; prior work is cited only bibliographically.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The supplemental material includes experiment code and run instructions. No new dataset or model is introduced.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology, theoretical development, or scientific claims of this research.

    Guidelines:

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