Efficient Thompson Sampling for Graph Bandits

Taro Iyadomi[1] Larissa Xu[2] Chenzhang Zhao[2] William Chang[2]

[1]Northwestern University [2]University of California, Los Angeles

Abstract

The multi-armed bandit (MAB) framework is used for sequential decision-making under uncertainty in applications with correlated or structured options. Prior work has used graph-based models to exploit dependencies between arms, reducing sample complexity in pure exploration tasks. Most graph-based methods to date use upper confidence bound (UCB) algorithms, while probabilistic sampling methods such as Thompson Sampling (TS) have not been analyzed in this setting. We develop TS algorithms for pure exploration in graph-structured bandits that combine graph-based information propagation with the tighter stopping criteria of randomized sampling. Our methods propagate information across connected arms using graph smoothness and attain sample complexity that depends on a graph-aware hardness rather than a sum over all arms. We provide sample-complexity guarantees, implementations for large graphs, and extensions to graph feedback, general PSD kernels, and misspecified graphs. The bounds reduce to the standard rate when the graph carries no information and improve on prior UCB-based bounds when it does.

1 Introduction

The multi-armed bandit (MAB) framework models sequential decision-making under uncertainty, with applications to clinical trials, recommendation systems, hyperparameter optimization, and online advertising [33]. The learner faces a trade-off between exploration and exploitation: it must collect enough information to identify the best option while limiting the number of samples. In the pure exploration variant, the objective is to identify the optimal arm (or one within ζ of the optimum) with confidence at least $1 - \delta$, rather than to maximize cumulative reward [6, 28].

When arms exhibit dependencies (as in recommendation systems, social networks, or drug discovery pipelines), these can be encoded as a graph of similarities or correlations [22, 37]. Graph-structured algorithms such as GRUB [47] use smoothness of rewards across connected nodes, so that observations of one arm inform estimates of its neighbors and the sample complexity scales with a graph-aware hardness rather than the number of arms. Existing graph-based algorithms in this setting use UCB-type confidence bounds, which incur a width-factor loss in structured domains [52].

Thompson Sampling (TS) [48] provides an alternative: by sampling from a distribution over arm rewards rather than maintaining deterministic confidence bounds, TS attains tighter sample complexity in structured exploration [52]. Existing TS analyses cover unstructured and combinatorial settings. The graph-structured case has not been treated.

We treat this case by developing TS algorithms for graph-structured pure exploration that combine graph-based information propagation with the stopping criteria of randomized sampling. We provide sample-complexity guarantees for four settings: graph-smooth bandits, misspecified graphs, graph feedback, and general PSD kernels. We then show that graph structure reduces the hardness from a sum over all arms to a smaller graph-aware quantity. Related work is discussed in Section A.

Our Contributions.

1. **Graph-smooth bandits (Section 3.1).** We adapt TS-Explore to the Laplacian-regularized setting, where the graph enters through a smoothness prior on the mean rewards. The resulting algorithm (Algorithm 1) achieves sample complexity $O\big(H_{\text{graph}} \log(1/\delta) + H_{\text{graph}} \log^2(K H_{\text{graph}})\big)$, where H_{graph} replaces the classical hardness $\sum_i \Delta_i^{-2}$ with a graph-aware quantity that separates competitive from non-competitive arms.

2. **Misspecified graphs (Section 3.2).** When the smoothness parameter ε is misspecified, the bias term in the confidence radius scales as $\sqrt{\rho}\,\varepsilon$ while the graph bonus grows linearly in ρ, so increasing ρ reduces the elimination threshold to a ρ-independent floor. We give an asymptotic analysis: as $\varepsilon \to 0$ at any $\rho \geq \rho_{\text{var}}(\varepsilon) := \sigma_0^2 L_1(T)/\varepsilon^2$, the competitive set decreases until $H_\varepsilon \to \max_{i \neq a^\star} \Delta_{i,c}^{-2}$, replacing the sum over competitive arms in H_{graph} with a maximum.

3. **Graph feedback (Section 3.3).** We consider the side-observation model in which pulling one arm reveals the rewards of all neighbors. The hardness measure H_{GF} is defined through a covering LP on the graph and is smaller than the graph-smooth hardness when the graph is dense.

4. **General PSD kernels (Section 3.4).** We generalize the Laplacian regularizer to an arbitrary positive semi-definite kernel K_G on the graph, recovering the combinatorial Laplacian as a special case. This is appropriate when the graph has non-uniform degrees and a normalized smoothness notion is preferred [15].

All proofs are deferred to the appendix (Appendices B–G).

2 Preliminaries

Let $[K] := \{1, 2, \ldots, K\}$ denote the set of arms, where each arm $i \in [K]$ is associated with an unknown reward distribution ν_i with mean $\mu_i = \mathbb{E}_{X \sim \nu_i}[X]$ and noise that is σ-sub-Gaussian, i.e. $\mathbb{E}[\exp(\lambda(X - \mu_i))] \leq \exp\big(\lambda^2\sigma^2/2\big)$ for all $\lambda \in \mathbb{R}$. (Bounded rewards on $[0, 1]$ are a special case with $\sigma = 1/2$.) At each round t, the learner chooses an arm $\pi_t \in [K]$ and observes a reward $r_{t,\pi_t} \sim \nu_{\pi_t}$ independently. The goal of the learner is to identify the optimal arm

$$a^\star = \arg\max_{i \in [K]} \mu_i$$

(or a ζ-optimal arm satisfying $\mu_i \geq \mu_{a^\star} - \zeta$) with probability at least $1 - \delta$, while minimizing the total number of arm pulls. This is the classical *fixed-confidence best-arm identification* setting [6, 28].

Graph representation of arms. Following [47], we assume the arms form the vertex set of a weighted undirected graph $G = (V_G, E_G, A_G)$ with $V_G = [K]$. The adjacency matrix $A_G \in \mathbb{R}^{K \times K}$ encodes similarities between arms, and the combinatorial Laplacian is $L_G = D_G - A_G$, where $D_G = \text{diag}(A_G \mathbf{1}_K)$. The reward vector $\mu \in \mathbb{R}^K$ is said to be *smooth on* G if the Laplacian seminorm

$$\|\mu\|_G \;=\; \sqrt{\langle \mu, L_G \mu \rangle} = \sqrt{\sum_{\{i,j\} \in E_G} A_{ij}\,(\mu_i - \mu_j)^2} \tag{1}$$

is small. Throughout this paper we assume $\|\mu\|_G \leq \varepsilon$. The smoothness assumption permits information transfer across neighbors, reducing the number of direct pulls per arm.

Notation. For a best arm $a^\star$, write $\Delta_i := \mu_{a^\star} - \mu_i \geq 0$ for the gap of arm $i \in [K]$ (with $\Delta_{a^\star} = 0$). The graph-structured analysis uses the *comparison gap* $\Delta_{i,c} := \Delta_i$ for $i \neq a^\star$. The subscript c emphasizes that this is the gap involved in the pairwise comparison test of the TS-Explore stopping rule, and is undefined for $i = a^\star$. Let $\Phi(x, \mu, \sigma^2)$ denote the probability that a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is at least x. For any $x \in (0, 0.5)$, define $\phi(x)$ by $\Phi(\phi(x), 0, 1) = x$. In the absence of graph side information, the sample complexity of best-arm identification scales as $\sum_{i \neq a^\star} \Delta_i^{-2}$ [6, 28]. Graph bandits improve on this by introducing the *effective number of plays* $t_{\text{eff},i}$, which aggregates direct samples and indirect information from neighbors, and the *influence factor* $\mathcal{J}(i, G)$ [54, 47], which quantifies the coupling between arms through the graph. These notions partition the suboptimal arms into a *competitive set* $\mathcal{H}$, requiring direct sampling, and a *non-competitive set*, which is eliminated by graph propagation.

3 Main Results

We present results for four settings, each using graph structure in a different way. In every case, the algorithm follows the TS-Explore template: draw $M(\delta, q, t)$ independent Thompson samples at each round, stop when all samples agree on the empirical best arm, and otherwise pull from the arm pair with the largest disagreement margin. The settings differ in the estimator, the effective sample count, and the hardness measure.

3.1 Graph-Smooth Bandits with Laplacian Regularization

Given a sequence of played arms $\pi_1, \ldots, \pi_T$ and observed rewards r_{t,π_t}, the Laplacian-regularized least-squares estimator is

$$\widehat{\boldsymbol{\mu}}_T = \left(\sum_{t=1}^{T} e_{\pi_t} e_{\pi_t}^\top + \rho L_G \right)^{-1} \left(\sum_{t=1}^{T} e_{\pi_t} r_{t,\pi_t} \right), \tag{2}$$

where $\rho > 0$ is a regularization parameter. The effective number of plays of arm i is

$$t_{\text{eff},i}(t) := \left[\left(\sum_{s=1}^{t} e_{\pi_s} e_{\pi_s}^\top + \rho L_G \right)^{-1} \right]_{ii}^{-1},$$

the graph-aware counterpart of the direct sample count [47, 1].

At each round t, define

$$M(\delta, q, t) := \left\lfloor \frac{1}{q} \log\left(\frac{12K^2 t^2}{\delta} \right) \right\rfloor, \qquad C(\delta, q, t) := \frac{\log\big(12K^2 t^2/\delta\big)}{\phi^2(q)}.$$

For each arm $i \in [K]$ and each sample index $m \in [M(\delta, q, t)]$, draw $\theta_i^m(t) \sim \mathcal{N}\big(\widehat{\mu}_i(t),\ C(\delta, q, t)/t_{\text{eff},i}(t)\big)$. If all sampled maximizers agree with the empirical maximizer, the algorithm stops and returns that arm. Otherwise it pulls the less-sampled arm among the empirical best and the sampled challenger with the largest disagreement margin. The full procedure is given in Algorithm 1.

The advantage of the agreement-based stopping rule over UCB–LCB elimination (as in CLUCB [13] or GRUB [47]) is that independent Thompson samples avoid the width-factor looseness of [52]: the disagreement test on samples gives a tighter elimination criterion than armwise confidence intervals. The mechanism extends to the graph-smooth setting once the direct pull count is replaced by $t_{\text{eff},i}(t)$, and the complexity then depends on the graph-aware hardness H_{graph} rather than a

Algorithm 1 `TS-Explore` for graph-structured pure exploration

Inputs: error constraint δ, threshold $q \in [\delta, 0.1]$, graph G, regularization parameter ρ.
Initialize direct counts $N_i \leftarrow 0$ and reward sums $R_i \leftarrow 0$ for all $i \in [K]$.
Pull one arm in each connected component of G (a total of C pulls) so that the regularized design matrix is invertible; set $t \leftarrow C$.
while true **do**
  Compute $V_t = \sum_{s=1}^{t} e_{\pi_s} e_{\pi_s}^\top + \rho L_G, \quad \widehat{\boldsymbol{\mu}}(t) = V_t^{-1} \sum_{s=1}^{t} e_{\pi_s} r_{s,\pi_s}$.
  For each arm i, set $t_{\mathrm{eff},i}(t) = \left[(V_t^{-1})_{ii}\right]^{-1}$.
  $\widehat{i}_t \leftarrow \arg\max_{i \in [K]} \widehat{\mu}_i(t)$.
  for $m = 1, \ldots, \lceil M(\delta, q, t) \rceil$ **do**
    for $i = 1, \ldots, K$ **do**
      Draw $\theta_i^m(t) \sim \mathcal{N}\big(\widehat{\mu}_i(t),\ C(\delta, q, t)/t_{\mathrm{eff},i}(t)\big)$.
    end for
    $\widetilde{i}_t^m \leftarrow \arg\max_{i \in [K]} \theta_i^m(t)$.
    $\widehat{\Delta}_t^m \leftarrow \theta^m_{\widetilde{i}_t^m}(t) - \theta^m_{\widehat{i}_t}(t)$.
  end for
  if $\widetilde{i}_t^m = \widehat{i}_t$ for all $m = 1, \ldots, \lceil M(\delta, q, t) \rceil$ **then**
    return $\widehat{i}_t$
  else
    $m^\star \leftarrow \arg\max_m \widehat{\Delta}_t^m$
    $\widetilde{i}_t \leftarrow \widetilde{i}_t^{m^\star}$
    Pull $\pi_{t+1} \in \arg\min_{i \in \{\widehat{i}_t, \widetilde{i}_t\}} t_{\mathrm{eff},i}(t)$, observe $r_{t+1,\pi_{t+1}}$, update data, and set $t \leftarrow t+1$.
  end if
end while

width-inflated quantity. A comparison with successive-elimination and top-two TS variants is given in Appendix B.

The analysis uses three good events: G_0 (concentration of the graph-regularized estimator), G_1 (Thompson samples close to their empirical means), and G_2 (for every pair of arms, at least one Thompson sample separates them by their true gap). The definitions and the proof that $\Pr(G_0 \cap G_1 \cap G_2) \geq 1 - \delta$ are given in Appendix C.

Lemma 1 $\Pr(G_0 \cap G_1 \cap G_2) \geq 1 - \delta$.

Theorem 2 (Correctness) *For any $q \in [\delta, 0.1]$, with probability at least $1 - \delta$, Algorithm 1 returns the optimal arm.*

Lemma 3 *Under the event $G_0 \cap G_1 \cap G_2$, a base arm i will not be pulled once $t_{\mathrm{eff},i}(t) \geq 186\, C(\delta, q, t)\, L(t)/\Delta_{i,c}^2$.*

To translate the elimination threshold into a horizon bound, we use the graph-induced lower bound on the effective sample size due to [47]:

Theorem 4 (Theorem C.1 of [47]) *Let π_T be the sampling policy up to time T. Then for every arm i,*

$$t_{\mathrm{eff},i} \ \geq \ t_i + \frac{1}{2} \left\lfloor \min \left\{ \rho\, \mathcal{J}(i, G),\ \sum_{j \in C(i)} t_j \right\} \right\rfloor, \tag{3}$$

where t_i is the number of direct pulls of arm i, $C(i)$ is its connected component, and $\mathcal{J}(i,G)$ is the influence factor.

Theorem 5 (Sample complexity in the graph-smooth model) *For $q \in [\delta, 0.1]$, with probability at least $1-\delta$, Algorithm 1 outputs the optimal arm with complexity upper bounded by*

$$O\left(H_{\text{graph}}\left(\log\frac{1}{\delta}+\log(KH_{\text{graph}})\right)\frac{\log(1/\delta)}{\log(1/q)}+H_{\text{graph}}\frac{\log^2(KH_{\text{graph}})}{\log(1/q)}\right),$$

where

$$H_{\text{graph}} \asymp \sum_{i\in\mathcal{H}\setminus\{a^\star\}}\frac{1}{\Delta_{i,c}^2}+\max_{i\in\mathcal{N}}\frac{1}{\Delta_{i,c}^2},$$

and $\mathcal{H},\mathcal{N}$ denote the competitive and non-competitive sets induced by the graph. In particular, choosing $q=\delta$ gives $O\big(H_{\text{graph}}\log\frac{1}{\delta}+H_{\text{graph}}\log^2(KH_{\text{graph}})\big)$.

The proof (Appendix C) follows the GRUB decomposition of [47]: non-competitive arms are eliminated quickly because their effective counts are augmented through the graph, while competitive arms determine the binding rate.

3.2 Misspecification and Smoothness Asymptotics

A concern is that the smoothness parameter ε may not be known. The naive expectation is a bias–variance trade-off in ρ, but the correct scaling gives a bias $\propto\sqrt{\rho}\,\varepsilon$, not $\rho\varepsilon$, so increasing ρ strictly improves the elimination threshold up to a ρ-independent floor. The confidence width supplied by the graph-regularized estimator (Lemma 3.2 of [47]) is

$$\gamma_i(t)=\frac{1}{\sqrt{t_{\text{eff},i}(t)}}\big(\underbrace{\sigma_0\sqrt{L_1(t)}}_{\text{variance}}+\underbrace{\sqrt{\rho}\,\varepsilon}_{\text{bias}}\big).$$

The $\sqrt{\rho}$ scaling follows from $V_t\succeq\rho L_G \Rightarrow V_t^{-1}\preceq\frac{1}{\rho}L_G^\dagger$ on range(L_G) (computation in Appendix D). The bias grows as $\sqrt{\rho}$ while the graph bonus $\rho\,\mathcal{J}(i,G)/2$ grows linearly, so there is no finite interior $\rho^\star$ that balances the two. An arm i is non-competitive when the graph bonus $t_{\text{eff},i}\geq\rho\,\mathcal{J}(i,G)/2$ alone exceeds the elimination threshold, which requires

$$\Delta_{i,c}\ \gtrsim\ \frac{\sigma_0\sqrt{L_1(t)}+\sqrt{\rho}\,\varepsilon}{\sqrt{\rho\,\mathcal{J}(i,G)/2}}\ =\ \sqrt{\tfrac{2}{\mathcal{J}(i,G)}}\left(\frac{\sigma_0\sqrt{L_1(t)}}{\sqrt{\rho}}+\varepsilon\right). \tag{4}$$

The right-hand side is strictly decreasing in ρ with limit $\varepsilon\sqrt{2/\mathcal{J}(i,G)}$, so the analysis recommends taking ρ at least large enough that the variance contribution is dominated by the bias floor:

$$\rho\ \geq\ \rho_{\text{var}}(\varepsilon)\ :=\ \frac{\sigma_0^2\,L_1(T)}{\varepsilon^2}. \tag{5}$$

The recommended scale is *quadratic* in $1/\varepsilon$, and above this threshold the elimination requirement is at most a constant multiple of $\varepsilon\sqrt{2/\mathcal{J}(i,G)}$. The $L_1(T)$ in (5) depends on the unknown stopping time, but only logarithmically, and a static estimate at the per-run pull budget suffices in practice (Appendix D).

Definition 6 (Competitive set in the large-ρ regime) *For a smoothness level $\varepsilon > 0$, define the competitive and non-competitive sets in the regime $\rho \geq \rho_{\mathrm{var}}(\varepsilon)$ as*

$$\mathcal{H}_\varepsilon := \Big\{ i \neq a^\star : \Delta_{i,c}^2 \, \mathcal{J}(i, G) \;\leq\; c_0 \, \varepsilon^2 \Big\},$$
$$\mathcal{N}_\varepsilon := [K] \setminus (\mathcal{H}_\varepsilon \cup \{a^\star\}),$$

where $c_0 > 0$ is an absolute constant from the analysis (derived as $c_0 = 8$ in Appendix D).

Theorem 7 (Sample complexity under recommended regularization) *Fix $q \in [\delta, 0.1]$. Suppose Algorithm 1 is run with $\rho \geq \rho_{\mathrm{var}}(\varepsilon)$. Then, with probability at least $1 - \delta$, the algorithm returns the optimal arm with stopping time*

$$O\left(H_\varepsilon \left(\log \frac{1}{\delta} + \log(K H_\varepsilon) \right) \frac{\log(1/\delta)}{\log(1/q)} + H_\varepsilon \frac{\log^2(K H_\varepsilon)}{\log(1/q)} \right),$$

where the smoothness-aware graph hardness is

$$H_\varepsilon \asymp \sum_{i \in \mathcal{H}_\varepsilon \setminus \{a^\star\}} \frac{1}{\Delta_{i,c}^2} + \max_{i \in \mathcal{N}_\varepsilon} \frac{1}{\Delta_{i,c}^2}.$$

Compared with Theorem 5, the fixed competitive set $\mathcal{H}$ is replaced by the ε-dependent set $\mathcal{H}_\varepsilon$. The threshold at which arm i leaves the competitive set is independent of ρ provided $\rho \geq \rho_{\mathrm{var}}(\varepsilon)$, since the $\sqrt{\rho}$-bias removes the bias–variance trade-off in ρ. Each arm i has a *critical smoothness level*

$$\varepsilon_i^\star := \Delta_{i,c} \sqrt{\frac{\mathcal{J}(i, G)}{c_0}}; \tag{6}$$

for $\varepsilon < \varepsilon_i^\star$, arm i exits $\mathcal{H}_\varepsilon$ into $\mathcal{N}_\varepsilon$ and is eliminated by graph propagation, with arms exiting in order of $\Delta_{i,c}^2 \, \mathcal{J}(i, G)$. The asymptotic limit is given by the following corollary.

Corollary 8 *Let $(\boldsymbol{\mu}^{(n)}, G^{(n)})_{n \geq 1}$ be a sequence of instances with the gaps $\Delta_{i,c}$ and the graph influence factors $\mathcal{J}(i, G^{(n)})$ held bounded away from zero, satisfying $\varepsilon_n := \|\boldsymbol{\mu}^{(n)}\|_{G^{(n)}} \to 0$. Run Algorithm 1 on instance n with $\rho \geq \rho_{\mathrm{var}}(\varepsilon_n)$. Then as $n \to \infty$:*

(i) $\mathcal{H}_{\varepsilon_n} \to \{a^\star\}$, i.e. every truly suboptimal arm becomes non-competitive.

(ii) $H_{\varepsilon_n} \to \max_{i \neq a^\star} \Delta_{i,c}^{-2}$, which is the minimum achievable hardness (matching the cost of verifying a single arm).

(iii) The sample-complexity bound of Theorem 7 approaches $O\big(\max_{i \neq a^\star} \Delta_{i,c}^{-2} \cdot \mathrm{polylog}(K, 1/\delta)\big)$.

3.3 Graph Feedback

We now consider a setting in which the graph determines the feedback structure: pulling one arm reveals the rewards of all arms in its closed neighborhood. This *graph-feedback* (or side-observation) model has been studied in regret-minimization settings [5, 53] and arises in applications such as social-network experiments and sensor placement. Let $G = (V_G, E_G)$ be an undirected graph with $V_G = [K]$ (so the graph has $n := K$ vertices, one per arm; we use n and K interchangeably in this section). If the learner pulls arm π_t, it observes rewards from all arms in $N_G^+(\pi_t) := N_G(\pi_t) \cup \{\pi_t\}$. The feedback observation count for arm i is $N_i^{\mathrm{fb}}(t) := \sum_{s=1}^t \mathbf{1}\{i \in N_G^+(\pi_s)\}$, and the estimator is the

empirical mean $\widehat{\mu}_i(t) = R_i^{\mathrm{fb}}(t)/N_i^{\mathrm{fb}}(t)$, which is unbiased with i.i.d. side observations, so standard sub-Gaussian concentration applies with no Laplacian bias term. Thompson samples are drawn as $\theta_i^k(t) \sim \mathcal{N}\big(\widehat{\mu}_i(t),\ C(\delta,q,t)/N_i^{\mathrm{fb}}(t)\big)$. Upon disagreement, the algorithm pulls the less-observed arm of the disagreement pair $\{\widehat{i}_t, \widetilde{i}_t\}$, and each such pull yields rewards from the entire closed neighborhood. The pseudocode is given as Algorithm 2 in Appendix E, and our implementation in Appendix H uses a more informative variant that selects $\arg\max_a |N_G^+(a) \cap \{\widehat{i}_t, \widetilde{i}_t\}|$, for which the same sample-complexity bound applies. The same template (Algorithm 1) and proof technique cover both the smooth and the feedback settings, with only the estimator and the candidate-cover step changing.

The graph feedback gain is captured by the allocation hardness

$$H_{\mathrm{GF}} = \min_{\boldsymbol{\tau} \in \mathbb{R}_+^n} \left\{ \sum_{a=1}^{n} \tau_a : \sum_{a:\, i \in N_G^+(a)} \tau_a \geq \frac{1}{\Delta_i^2} \text{ for all } i \neq a^\star \right\}. \tag{7}$$

This is a covering LP: it minimizes the number of pulls subject to the constraint that every suboptimal arm i accumulates at least Δ_i^{-2} observations through the neighborhoods of pulled arms. The LP is solvable in polynomial time, and Corollary 13 gives explicit bounds in terms of standard graph parameters.

Lemma 9 $\Pr\big(G_0^{\mathrm{fb}} \cap G_1^{\mathrm{fb}} \cap G_2^{\mathrm{fb}}\big) \geq 1 - \delta$.

Theorem 10 (Correctness) *For any $q \in [\delta, 0.1]$, with probability at least $1-\delta$, Algorithm 2 returns the optimal arm.*

Lemma 11 *Under $G_0^{\mathrm{fb}} \cap G_1^{\mathrm{fb}} \cap G_2^{\mathrm{fb}}$, arm i will not trigger further exploration once $N_i^{\mathrm{fb}}(t) \geq 186\, C(\delta,q,t)\, L_2(t)/\Delta_i^2$.*

Theorem 12 (Sample complexity under graph feedback) *For $q \in [\delta, 0.1]$, with probability at least $1-\delta$, Algorithm 2 outputs the optimal arm with complexity*

$$O\left(H_{\mathrm{GF}} \log n \left(\log \frac{1}{\delta} + \log(nH_{\mathrm{GF}}) \right) \frac{\log(1/\delta)}{\log(1/q)} + H_{\mathrm{GF}} \log n \, \frac{\log^2(nH_{\mathrm{GF}})}{\log(1/q)} \right).$$

In particular, if $q = \delta$, then $O\big(H_{\mathrm{GF}} \log n \log \frac{1}{\delta} + H_{\mathrm{GF}} \log n \log^2(nH_{\mathrm{GF}})\big)$. The $\log n$ factor is the standard set-cover integrality gap arising from restricting the pull rule to the disagreement pair $\{\widehat{i}_t, \widetilde{i}_t\}$ rather than the full action set $[n]$, and is removable on instances where the disagreement candidates already cover the elimination requirement (e.g. regular graphs with uniform gaps).

Corollary 13 (Combinatorial bounds on H_{GF}) *Let $\Delta_{\min} := \min_{i \neq a^\star} \Delta_i$ and $\Delta_{\max} := \max_{i \neq a^\star} \Delta_i$. Let $d_{\min}(G) := \min_{v \in [n]} \deg_G(v)$ be the minimum vertex degree, $\bar{\chi}(G) := \chi(\bar{G})$ the clique-cover number of G, and $\rho(G)$ the 2-packing number of G (the size of the largest set of vertices whose closed neighborhoods are pairwise disjoint). Then*

$$(\rho(G) - 1) \cdot \Delta_{\min}^{-2} \ \leq\ H_{\mathrm{GF}} \ \leq\ \bar{\chi}(G) \cdot \Delta_{\max}^{-2} \ \leq\ \frac{n}{1 + d_{\min}(G)}\, \Delta_{\max}^{-2}. \tag{8}$$

The -1 in the lower bound accounts for the case in which a maximum 2-packing of G contains the optimal arm $a^\star$, which contributes no constraint to the LP (7). If $a^\star$ is not in any maximum 2-packing, the bound improves to $\rho(G)\Delta_{\min}^{-2}$. Moreover, when $\Delta_i \equiv \Delta$ for all $i \neq a^\star$, $H_{\mathrm{GF}} = \Delta^{-2}\rho^*(G)$ by LP duality, where $\rho^*(G)$ is the fractional 2-packing number.

The clique-cover bound expresses the improvement from feedback (few cliques imply few pulls suffice), while the 2-packing lower bound expresses the obstruction from disjoint neighborhoods. Closed-form evaluations on canonical graph families (empty, complete, star, d-regular) are given in Appendix E. Compared with Theorem 5, both bounds share the form $O(H\log(1/\delta) + H\log^2(\cdot))$, but H_{graph} uses *spectral propagation* through the Laplacian-regularized estimator, whereas H_{GF} uses *combinatorial covering* via side observations: H_{GF} interpolates from $\sum_i \Delta_i^{-2}$ on the empty graph to Δ^{-2} on the clique, and is $\Theta(\max_i \Delta_i^{-2})$ for dense graphs (potentially smaller than H_{graph}). The proofs are in Appendix E and follow the graph-smooth analysis after replacing $t_{\text{eff},i}(t)$ with $N_i^{\text{fb}}(t)$ and using the feedback LP to convert per-arm thresholds into a total pull count.

3.4 General PSD Graph Kernels

When the graph has highly non-uniform degrees, the combinatorial Laplacian L_G can be dominated by high-degree vertices. One alternative is to replace L_G with the *normalized Laplacian* $K_G := I - D_G^{-1/2} A_G D_G^{-1/2}$ [15], or more generally any PSD kernel. We assume $\langle \boldsymbol{\mu}, K_G \boldsymbol{\mu} \rangle \le \varepsilon^2$ and use the regularized estimator $\widehat{\boldsymbol{\mu}}_T = (V_T)^{-1} \sum_{t=1}^T e_{\pi_t} r_{t,\pi_t}$ with $V_T = \sum_{t=1}^T e_{\pi_t} e_{\pi_t}^\top + \rho K_G$ and kernel-effective sample sizes $t_{\text{eff},i}(t) := [(V_t)^{-1}]_{ii}^{-1}$. The algorithm is identical to Algorithm 1 with L_G replaced by K_G. Pseudocode is deferred to Algorithm 3 in Appendix F.

Theorem 14 *Fix $q \in [\delta, 0.1]$. Assume $\langle \boldsymbol{\mu}, K_G \boldsymbol{\mu} \rangle \le \varepsilon^2$. Then with probability at least $1-\delta$, Algorithm 3 returns the optimal arm $a^\star$. Moreover, its stopping time is upper bounded by the smallest T such that*

$$t_{\text{eff},i}(T) \;\ge\; \frac{196\, C(T)\, L(T)}{\Delta_{i,c}^2} \qquad \text{for all } i \neq a^\star.$$

The sample complexity has the same gap-dependent form as Theorem 5, with effective sample sizes computed using V_T.

The effective sample size admits the variational representation $t_{\text{eff},i}(T) = \min_{x \in \mathbb{R}^n,\, x_i = 1} \{ \sum_j t_j\, x_j^2 + \rho\, \langle x, K_G\, x \rangle \}$, from which we derive the following spectral bounds.

Proposition 15 (Spectral sandwich for t_{eff}) *Let $K_G \succeq 0$ have $\ker(K_G) = \operatorname{span}(v_0)$ for some unit vector $v_0 \in \mathbb{R}^n$, with nonzero eigenvalues $0 < \lambda_2(K_G) \le \cdots \le \lambda_n(K_G)$. For any sampling counts $\{t_j\}_{j=1}^n$ with $T = \sum_j t_j$ and every arm $i \in [n]$,*

$$t_i + \rho\lambda_2(K_G)\big(1 - (v_0)_i^2 - \varepsilon_T\big) \;\le\; t_{\text{eff},i}(T) \;\le\; t_i + \rho\lambda_n(K_G),$$

where $\varepsilon_T \to 0$ once every arm has been sampled $\Omega(\rho\lambda_2(K_G))$ times. For the combinatorial Laplacian L_G, $v_0 = \mathbf{1}/\sqrt{n}$ and $(v_0)_i^2 = 1/n$ uniformly. For the normalized Laplacian $\mathcal{L}_G = I - D_G^{-1/2} A_G D_G^{-1/2}$, $\ker(\mathcal{L}_G) = \operatorname{span}(D_G^{1/2}\mathbf{1})$, so $v_0 = D_G^{1/2}\mathbf{1}/\|D_G^{1/2}\mathbf{1}\|$ and $(v_0)_i^2 = \deg_G(i)/(2|E_G|)$ varies with the degree.

The spectral gap $\lambda_2(K_G)$ controls the rate at which unsampled arms acquire information, and $\lambda_n(K_G)$ bounds the regularizer's contribution. The condition number $\kappa(K_G) := \lambda_n/\lambda_2$ determines the uniformity of effective counts across arms. Setting $K_G = L_G$ recovers Theorem 5. The normalized Laplacian satisfies $\lambda_n(\mathcal{L}_G) \le 2$ universally, so $\kappa(\mathcal{L}_G) = O(1)$ on graphs where $\kappa(L_G) \asymp n$ (e.g., stars), yielding more uniform effective counts on degree-heterogeneous graphs. The full comparison is in Appendix F.

4 Experiments

The experiments in this section have two goals. First, we want to see whether the rate improvements predicted by Theorems 5 and 12 are empirically visible on synthetic and real graphs. Second, we want to compare the resulting algorithms against natural UCB-based and graph-free baselines. We present two figures: one for the graph-smooth setting on a synthetic two-cluster instance and on the MovieLens-100K recommender graph, and one for the graph-feedback setting on Erdős–Rényi graphs of varying density. All experiments use $\delta = 10^{-3}$, $q = 0.1$, and Gaussian rewards with unit variance. We report medians and 25–75 interquartile ranges (IQR) over 20 seeds, and every run returned the optimal arm. Two further studies in Appendix H test the smoothness-asymptotics result of Corollary 8 and the normalized-kernel result of Theorem 14. Full protocol details are also in the appendix.

Graph-smooth pure exploration (Figure 1). We first test Theorem 5 on a controlled synthetic instance, then on a real-world recommender graph. The synthetic instance is a two-cluster chain: a singleton best arm at $\mu_{a^\star} = 1$ is joined to a challenger clique whose $K - 1$ vertices share $\mu_i = 1 - \Delta$, with $\Delta = 0.3$ and $\rho = 100$. This is the prototypical setting that the graph-smooth analysis is designed for, since the $K - 1$ challengers form a single cluster on which the regularized estimator can pool observations. Sweeping K from 20 to 200, the median stopping time of TS-Explore *decreases* with K, while Basic TS [52], KL-LUCB [29], and GRUB [47] all grow roughly linearly. The gap between TS-Explore and the fastest unstructured baseline (Basic TS) widens from a factor of roughly two at $K = 20$ to more than an order of magnitude at $K = 100$. This is the qualitative signature of the bound in Theorem 5: as the cluster grows, the per-cluster hardness in H_{graph} saturates while the unstructured hardness $\sum_i \Delta_i^{-2}$ continues to grow, so the gap between the two rates widens with K rather than shrinks.

The MovieLens-100K experiment then asks whether the same rate improvement is visible on a graph that arises from data rather than from construction. We take the $K = 20$ top-rated movies, build a top-5 mutual-neighbor item-item similarity graph (realized smoothness $\varepsilon \approx 3.41$, $\Delta_{\min} \approx 0.07$), and sweep ρ from 1 to 10^3. TS-Explore matches Basic TS at small ρ, where the graph regularizer contributes little. At $\rho = 10^3$, the predicted operating regime $\rho \geq \rho_{\text{var}}(\varepsilon)$ from Eq. (5) is reached, and the median stopping time of TS-Explore falls well below those of Basic TS and KL-LUCB. GRUB stays slower than the unstructured baselines across the swept ρ range. We do not attempt to match its empirical rate to its analytical bound on this instance. Together, the synthetic and real-data panels indicate that the rate improvement of Theorem 5 is not an artifact of stylized topology and survives on a graph with non-uniform degrees and tightly bunched gaps.

Graph feedback (Figure 2). We test Theorem 12 on Erdős–Rényi random graphs $G(n, p)$ with $n = 20$ vertices and uniform $\Delta = 0.3$, sweeping the edge density p from 0.05 to 1.0 over 20 realizations per setting. As p grows, the median stopping times of TS-Explore-GF and UCB+cover (a UCB–LCB analogue using the same candidate-pair-cover pull rule) both fall by roughly an order of magnitude. Both feedback algorithms drop below KL-LUCB [29] once p exceeds 0.1. The no-feedback references, TS-Explore at $\rho = 1$ and Basic TS, are constant in p, as expected: their stopping rules cannot exploit the side observations. This pattern is the qualitative signature predicted by Theorem 12, since the covering hardness H_{GF} shrinks faster than the unstructured hardness $\sum_i \Delta_i^{-2}$ as the graph densifies.

UCB+cover is in fact slightly faster than TS-Explore-GF across the swept range. This reflects a real difference between the two settings. In the graph-smooth setting (Figure 1), the

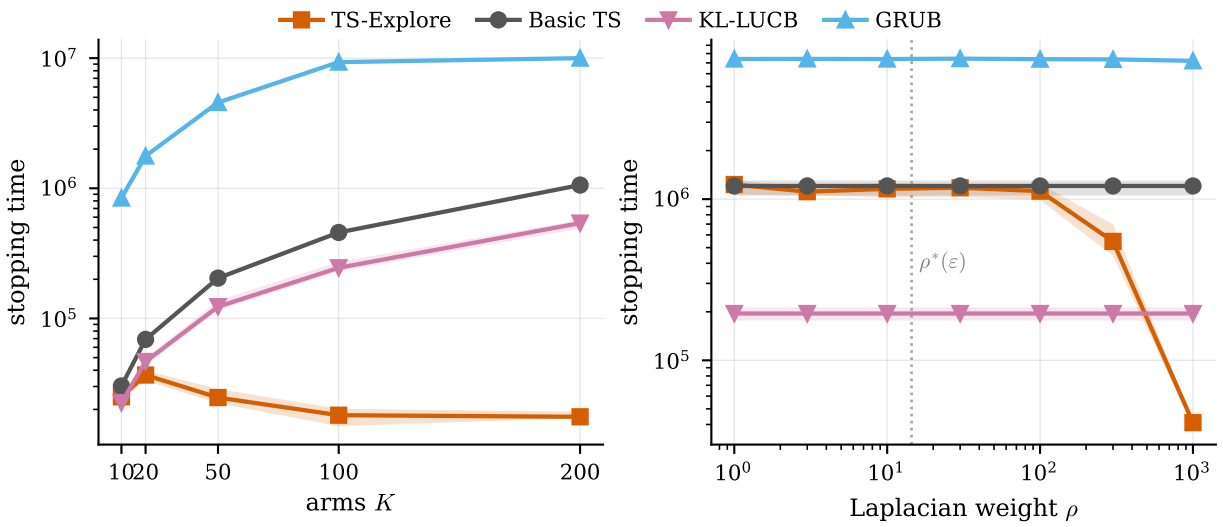


Figure 1: Graph-smooth pure exploration. **Left:** synthetic two-cluster chain, $K \in \{10, \ldots, 200\}$, $\Delta = 0.3$, $\rho = 100$. **Right:** MovieLens-100K, $K = 20$, top-5 mutual-neighbor similarity ($\varepsilon \approx 3.41$, $\Delta_{\min} \approx 0.07$). The dotted line marks $\rho_{\text{var}}(\varepsilon)$ from Eq. (5). Median and 25–75 IQR over 20 seeds.

regularizer introduces a bias term that inflates the UCB–LCB confidence radius, and the agreement-based TS stopping rule avoids that inflation. This is the width-factor advantage that Theorem 5 exploits. In the graph-feedback setting, the empirical-mean estimator is unbiased, so UCB–LCB has no width-factor handicap and can be the faster stopping rule on a given pull sequence. We extend TS-Explore to the feedback setting for *algorithmic unification*: a single stopping-rule template covers both settings, the proof structure is identical (only the estimator and concentration inequality change), and the resulting bound has the same dependence on a graph-aware hardness measure. The right panel of Figure 2 confirms this picture by separating the two design axes that distinguish TS-Explore-GF from a classical UCB-style feedback algorithm: the stopping rule (TS agreement versus UCB–LCB elimination) and the pull rule (candidate-pair cover versus argmax confidence width). Within each stopping rule, the cover-pair pull rule is faster than the width-based rule. Within each pull rule, UCB–LCB elimination is faster than TS agreement. The two effects are approximately additive.

Further studies in the appendix. Two appendix experiments check the remaining theoretical claims. Appendix H.3 tests the smoothness-asymptotics result of Corollary 8 on a connected stochastic block model. The stopping time of TS-Explore run with $\rho \geq \rho_{\text{var}}(\varepsilon)$ decreases monotonically as $\varepsilon \to 0$, while the un-tuned baselines stay constant at the no-graph rate, matching the qualitative shape that the corollary predicts. Appendix H.2 tests the normalized-kernel result of Theorem 14 on a Barabási–Albert hub-optimal instance. The normalized Laplacian K_G degrades more slowly than the combinatorial Laplacian L_G at large ρ, consistent with the spectral sandwich of Proposition 15.

5 Conclusion

We developed Thompson Sampling algorithms for pure exploration in graph-structured bandits and gave sample-complexity guarantees in four settings: graph-smooth bandits with Laplacian regularization, misspecified smoothness, graph feedback, and general PSD kernels. The bounds

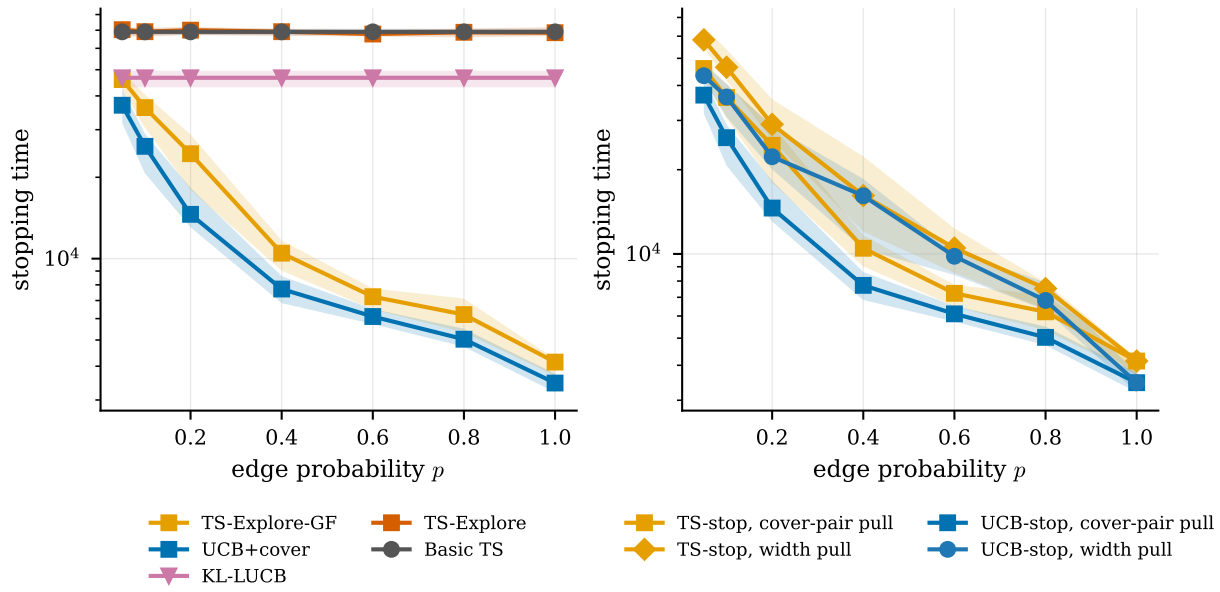


Figure 2: Graph feedback on Erdős–Rényi $G(20, p)$ with uniform $\Delta = 0.3$. **Left:** stopping time as a function of edge density p for two feedback algorithms and three baselines. **Right:** the four corners of the $\{\mathrm{TS}, \mathrm{UCB}\} \times \{\mathrm{cover}, \mathrm{width}\}$ design (color encodes the stopping rule, marker encodes the pull rule). Median and 25–75 IQR over 20 realizations per density.

replace the classical hardness $\sum_i \Delta_i^{-2}$ with a graph-aware quantity, recovering the standard rate when the graph carries no information and improving on prior UCB-based bounds when it does. The analysis identifies a $\sqrt{\rho}$ scaling of the bias term that removes the bias–variance trade-off in the regularization parameter and yields an explicit prescription $\rho \geq \rho_{\mathrm{var}}(\varepsilon)$. Open directions include a matching lower bound for H_{graph}, removal of the $\log n$ factor in the graph-feedback bound through an LP-rounded pull rule, and extension to the contextual setting.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[2] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, 2013.

[3] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[4] R. K. Ando and T. Zhang. Learning on graph with laplacian regularization. In *Advances in Neural Information Processing Systems*, volume 19, page 25, 2007.

[5] A. Atsidakou, O. Papadigenopoulos, C. Caramanis, S. Sanghavi, and S. Shakkottai. Asymptotically-optimal gaussian bandits with side observations. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[6] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory (COLT)*, pages 41–53, 2010.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[8] R. B. Bapat and S. Gupta. Resistance distance in wheels and fans. *Indian Journal of Pure and Applied Mathematics*, 41(1):1–13, 2010.

[9] D. A. Berry and B. Fristedt. *Bandit problems: sequential allocation of experiments*. Monographs on statistics and applied probability. Springer, 1985.

[10] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 23–37. Springer, 2009.

[11] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. *arXiv preprint arXiv:1210.4839*, 2012.

[12] L. Chen, A. Gupta, J. Li, M. Qiao, and R. Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pages 482–534, 2017.

[13] S. Chen, T. Li, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014.

[14] W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework, results and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

[15] F. R. Chung and F. C. Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

[16] R. Combes, M. Talebi, A. Proutiere, and M. Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, 2015.

[17] G. Dasarathy, R. Nowak, and X. Zhu. S2: An efficient graph based active learning algorithm with application to nonparametric classification. In *Conference on Learning Theory*, pages 503–522. PMLR, 2015.

[18] E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(Jun):1079–1105, 2006.

[19] V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[20] V. Gabillon, A. Lazaric, M. Ghavamzadeh, R. Ortner, and P. Bartlett. Improved learning complexity in combinatorial pure exploration bandits. In *Artificial Intelligence and Statistics*, pages 1004–1012, 2016.

[21] A. Garivier and E. Kaufmann. Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models. *arXiv preprint arXiv:1902.00671*, 2019.

[22] C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *International Conference on Machine Learning (ICML)*, 2014.

[23] S. Gupta, G. Joshi, and O. Yağan. Correlated multi-armed bandits with a latent random source. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3572–3576. IEEE, 2020.

[24] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[25] K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE, 2014.

[26] M. Jourdan, M. Mutny, J. Kirschner, and A. Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 805–849, 2021.

[27] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, 2012.

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

[29] E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251, 2013.

[30] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Conference on Algorithmic Learning Theory*, 2012.

[31] T. Kocák and A. Garivier. Best arm identification in spectral bandits. *arXiv preprint arXiv:2005.09841*, 2020.

[32] N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26, 2013.

[33] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[34] J. Lee, J. Honda, and M. Sugiyama. Thompson exploration with best challenger rule in best arm identification. In *Asian Conference on Machine Learning*, pages 646–661. PMLR, 2024.

[35] D. LeJeune, G. Dasarathy, and R. Baraniuk. Thresholding graph bandits with grapl. In *International Conference on Artificial Intelligence and Statistics*, pages 2476–2485. PMLR, 2020.

[36] J. Li, C. Du, and J. Zhu. A bayesian approach for subset selection in contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8384–8391, 2021.

[37] S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016.

[38] J. Lipor and G. Dasarathy. Quantile search with time-varying search parameter. In *Asilomar Conference on Signals, Systems, and Computers*, 2018.

[39] Y. Ma, R. Garnett, and J. Schneider. σ-optimality for active learning on gaussian random fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[40] Y. Ma, T.-K. Huang, and J. Schneider. Active search and bandits on graphs using sigma-optimality. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

[41] P. Perrault, E. Boursier, V. Perchet, and M. Valko. Statistical efficiency of thompson sampling for combinatorial semi-bandits. In *Advances in Neural Information Processing Systems*, 2020.

[42] N. Rao, H.-F. Yu, P. Ravikumar, and I. S. Dhillon. Collaborative filtering with graph information: Consistency and scalable methods. In *Advances in Neural Information Processing Systems*, volume 2, page 7, 2015.

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

[44] D. Russo. Simple bayesian algorithms for best arm identification. In *Conference on learning theory*, pages 1417–1418. PMLR, 2016.

[45] O. Shamir. A variant of azuma's inequality for martingales with subgaussian tails, 2011.

[46] X. Shang, R. Heide, P. Menard, E. Kaufmann, and M. Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832, 2020.

[47] P. Thaker, M. Malu, N. Rao, and G. Dasarathy. Maximizing and satisficing in multi-armed bandits with graph information. *Advances in Neural Information Processing Systems*, 35:2019–2032, 2022.

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

[49] M. Valko, R. Munos, B. Kveton, and T. Kocák. Spectral bandits for smooth graph functions. In *International conference on machine learning*, pages 46–54. PMLR, 2014.

[50] D. Wang, J. Lipor, and G. Dasarathy. Distance-penalized active learning via markov decision processes. In *2019 IEEE Data Science Workshop (DSW)*, pages 155–159. IEEE, 2019.

[51] S. Wang and W. Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122. PMLR, 2018.

[52] S. Wang and J. Zhu. Thompson sampling for (combinatorial) pure exploration. In *International Conference on Machine Learning*, pages 23470–23483. PMLR, 2022.

[53] Y. Wu, A. György, and C. Szepesvári. Online learning with gaussian payoffs and side observations, 2015.

[54] W. Xiao and I. Gutman. Resistance distance and laplacian spectrum. *Theoretical Chemistry Accounts*, 110(4):284–289, 2003.

[55] K. Yang, L. Toni, and X. Dong. Laplacian-regularized graph bandits: Algorithms and theoretical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 3133–3143. PMLR, 2020.

[56] Y. Zhou, X. Chen, and J. Li. Optimal pac multiple arm identification with applications to crowdsourcing. In *International Conference on Machine Learning*, pages 217–225. PMLR, 2014.

[57] X. J. Zhu. Semi-supervised learning literature survey, 2005.

A Related Work

Foundations of multi-armed bandits. The study of multi-armed bandits originates with [43], who introduced the sequential design of experiments with adaptive allocation. Subsequent classical treatments such as [9] gave a unified account of bandit problems and motivated subsequent work on adaptive experimentation. The finite-time analysis of [7] established regret guarantees for UCB and made it a standard algorithm in stochastic bandits. A modern treatment of the field, including both regret minimization and pure exploration, is given in [33].

Pure exploration and best-arm identification. The pure exploration setting has been studied extensively. [6] introduced UCB-E and Successive Rejects, with sample-complexity guarantees within constants of the lower bound. [10] introduced the notion of simple regret and gave trade-offs between exploration and recommendation accuracy. The PAC framework of [18] introduced action-elimination techniques applicable to reinforcement learning. Extensions such as LUCB and lil'UCB [27, 25] combined elimination and confidence-bound approaches. Information-theoretic lower bounds on best-arm identification are derived in [29, 28], with distinctions between the fixed-confidence and fixed-budget regimes. Additional refinements include improved stopping rules [21] and unified budget/confidence frameworks [19].

Combinatorial pure exploration. Combinatorial bandits extend the standard setting to structured action spaces. [14] introduced a general framework for combinatorial bandits with semi-bandit feedback, and [13] introduced the CLUCB algorithm for combinatorial pure exploration. [12] gave algorithms with sample complexity within constants of the lower bound for structured exploration, and [20] studied complexity bounds under combinatorial constraints. Online and scalable algorithms are proposed in [16, 26]. Related subset-selection problems are studied in [29] in terms of information complexity and in [56] for PAC multiple-arm identification. Quantile and thresholding formulations are addressed in [38, 35].

Graph-structured bandits and side information. A line of work uses graph-structured side information, where rewards are assumed smooth over a known graph. The spectral bandit framework of [49] showed that regret can scale with the effective dimension of the graph Laplacian rather than the number of arms. Extensions include best-arm identification in spectral bandits by [31]. Laplacian-regularized approaches such as GraphUCB [55] use user graphs for scalable recommendation and tighter regret bounds. Other applications include graph-regularized thresholding bandits [35], correlated bandits with latent sources [23], and graph-based active learning [39, 40, 50, 17]. The GRUB algorithm of [47] introduced the influence factor to describe how graph topology accelerates elimination of suboptimal arms. Side observations in bandits have also been studied in [5] and [53] in regret-minimization settings. Related work on graph-based semi-supervised learning [57], spectral graph theory [15], Laplacian regularization [4], resistance distances [54, 8], and smoothness priors [45, 42] informs the design of graph-structured bandit algorithms.

Thompson Sampling. Thompson Sampling (TS), originally introduced by [48], is a Bayesian probability-matching strategy for the exploration–exploitation trade-off. Subsequent analyses gave regret guarantees within constants of the lower bound: [3] proved regret bounds for linear contextual bandits, [32] extended the analysis to exponential families, and [2] and [30] refined the finite-time guarantees. In the combinatorial setting, [51] introduced CTS for semi-bandits, and [41] analyzed the statistical efficiency of TS for combinatorial semi-bandits. For pure exploration, [44] proposed Bayesian top-two sampling algorithms, [46] extended the Bayesian analysis to Gaussian settings with fixed-confidence guarantees, [34] proposed BC-TE, and [36] introduced Bayesian Re-sample Exploration for contextual bandits. The closest prior work to ours is [52], which proposed TS-Explore, the first Thompson Sampling algorithm for combinatorial pure exploration in the frequentist setting, with width-factor improvements over CLUCB.

B Discussion of Algorithm Design Choices

The pure-exploration character of Algorithm 1 is encoded in its stopping rule rather than its pull rule: termination requires that all $M(\delta, q, t)$ independent Thompson samples agree with the empirical maximizer, which, together with the choice of M and C, gives a high-probability certificate of optimality. The pull rule serves as the exploration step that produces this certificate as quickly as possible, with no regret-minimization objective. This is the same separation of stopping from sampling as in LUCB-style stopping rules used in pure exploration, but with random samples in place of confidence intervals.

Our reason for following the TS-Explore template of [52] rather than a successive-elimination scheme is the same as the motivation in the combinatorial pure-exploration setting they study. Successive-elimination and LUCB-style algorithms (e.g., CLUCB of [13] for combinatorial pure exploration or the GRUB algorithm of [47] in the graph-structured case) eliminate arms by comparing armwise upper and lower confidence bounds. In structured settings, the relevant uncertainty involves an aggregate of armwise quantities, and a sum of confidence radii is larger than the standard deviation of the corresponding sum of empirical means. This looseness is the source of the width-factor inefficiency in CLUCB identified in [52]. Independent Thompson samples avoid this slack: with high probability the sum of samples within a candidate set does not exceed the aggregate confidence bound, so the disagreement test on samples gives a tighter elimination criterion than armwise UCB/LCB intervals. The same mechanism extends to the graph-smooth setting once the direct pull count is replaced by $t_{\mathrm{eff},i}(t)$, and the complexity bound therefore depends on H_{graph} rather than on a width-inflated UCB-style quantity.

Top-two Thompson sampling variants such as [44] and [46] also use random sampling but make decisions from posterior distributions rather than realized samples, which leaves no direct stopping rule under the frequentist guarantees we target and is difficult to combine with graph regularization.

C Proofs for the Graph-Smooth Laplacian Model

In this appendix we give a detailed proof of the graph-smooth result, drawing on the GRUB analysis of [47] and the TS-Explore analysis of [52]. The argument shows that the Thompson-sampling rule inherits the same elimination structure as GRUB once the graph-regularized estimator is paired with the effective sample size $t_{\mathrm{eff},i}(t)$.

For convenience, write

$$L_1(t) := \log\left(\frac{12K^2t^2}{\delta}\right), \qquad L_2(t) := \log\left(\frac{12K^2t^2M(\delta,q,t)}{\delta}\right), \qquad C(t) := C(\delta,q,t).$$

We also let

$$\gamma_i(t) := \sqrt{\frac{1}{t_{\mathrm{eff},i}(t)}} \left(2\sigma \sqrt{14 \log\left(\frac{2w_i(\boldsymbol{\pi}_t)}{\delta}\right)} + \sqrt{\rho}\, \|\boldsymbol{\mu}\|_G \right),$$

which is the confidence width supplied by Lemma A.2 of [47] after applying $V_t^{-1} \preceq \frac{1}{\rho} L_G^{\dagger}$ on range(L_G). Since $\|\boldsymbol{\mu}\|_G \le \varepsilon$, we will repeatedly use the deterministic upper bound

$$\gamma_i(t) \le \sqrt{\frac{1}{t_{\mathrm{eff},i}(t)}} \left(2\sigma \sqrt{14 \log\left(\frac{2w_i(\boldsymbol{\pi}_t)}{\delta}\right)} + \sqrt{\rho}\, \varepsilon \right).$$

Finally, recall that $\Delta_{i,c} = \mu_{a^\star} - \mu_i$ for each suboptimal arm i.

Definition of the good events. We work on the intersection of the following three events.

$$G_0 := \Big\{ \forall t \ge 1,\ \forall i \in [K],\ |\widehat{\mu}_i(t) - \mu_i| \le \gamma_i(t) \Big\},$$

$$G_1 := \Big\{ \forall t \ge 1,\ \forall m \in [M(\delta, q, t)],\ \forall i \in [K],\ |\theta_i^m(t) - \widehat{\mu}_i(t)| \le \sqrt{\frac{2C(t)L_2(t)}{t_{\mathrm{eff},i}(t)}} \Big\},$$

$$G_2 := \Big\{ \forall t \ge 1,\ \forall i \ne j,\ \exists m \in [M(\delta, q, t)] \text{ such that } \theta_i^m(t) - \theta_j^m(t) \ge \mu_i - \mu_j \Big\}.$$

Event G_0 controls the graph-regularized estimator around the true mean vector, G_1 controls the Gaussian Thompson perturbations around the empirical means, and G_2 guarantees that among the $M(\delta, q, t)$ sample sets there is always at least one sample that separates any pair of arms by at least its true gap.

Proof of Lemma 1. We prove the three pieces separately.

Proof: We first bound the failure probability of G_0. By Lemma A.2 of [47], for every fixed time t and arm i,

$$\Pr(|\widehat{\mu}_i(t) - \mu_i| > \gamma_i(t)) \le \frac{\delta}{2w_i(\boldsymbol{\pi}_t)}.$$

The weights $w_i(\boldsymbol{\pi}_t)$ are chosen in that lemma so that the resulting tail probabilities are summable over all arms and all times. Consequently,

$$\Pr(G_0^c) \le \frac{\delta}{3},$$

after fixing the numerical constant hidden in the definition of $w_i(\boldsymbol{\pi}_t)$ exactly as in the proof of Lemma A.2 in [47].

We next control G_1. Fix $t \ge 1$, $m \in [M(\delta, q, t)]$, and $i \in [K]$. Conditional on the history up to time t, we have

$$\theta_i^m(t) - \widehat{\mu}_i(t) \sim \mathcal{N}\left(0, \frac{C(t)}{t_{\mathrm{eff},i}(t)}\right).$$

Therefore, by standard Gaussian concentration,

$$\Pr\left[|\theta_i^m(t) - \widehat{\mu}_i(t)| > \sqrt{\frac{2C(t)L_2(t)}{t_{\mathrm{eff},i}(t)}} \right] \le 2e^{-L_2(t)}.$$

Since

$$2e^{-L_2(t)} = \frac{\delta}{6K^2t^2M(\delta,q,t)},$$

a union bound over arms, Thompson samples, and times yields

$$\Pr(G_1^c) \le \sum_{t\ge1}\sum_{m=1}^{M(\delta,q,t)}\sum_{i=1}^{K}\frac{\delta}{6K^2t^2M(\delta,q,t)} = \sum_{t\ge1}\frac{\delta}{6Kt^2} \le \sum_{t\ge1}\frac{\delta}{6t^2} \le \frac{\delta\,\pi^2}{36} < \frac{\delta}{3}.$$

[1]

We now prove the G_2 bound. Fix a time t and two distinct arms i, j. Conditional on the empirical means,

$$\theta_i^m(t) - \theta_j^m(t) \sim \mathcal{N}\Big(\widehat{\mu}_i(t) - \widehat{\mu}_j(t),\ C(t)\Big(\frac{1}{t_{\mathrm{eff},i}(t)} + \frac{1}{t_{\mathrm{eff},j}(t)}\Big)\Big).$$

On G_0 we have

$$\widehat{\mu}_i(t) - \widehat{\mu}_j(t) \ge \mu_i - \mu_j - \gamma_i(t) - \gamma_j(t).$$

Let $\Phi(\cdot,0,1)$ denote the standard-normal upper-tail (survival) function from Section 4. Since the distribution has standard deviation $\sqrt{C(t)(1/t_{\mathrm{eff},i}+1/t_{\mathrm{eff},j})}$ and the mean is shifted from $\mu_i-\mu_j$ by at most $-\gamma_i(t)-\gamma_j(t)$, we obtain, conditional on G_0,

$$\begin{aligned}&\Pr\big[\theta_i^m(t) - \theta_j^m(t) \ge \mu_i - \mu_j \mid G_0\big]\\ &\quad\ge \Phi\left(\frac{\gamma_i(t)+\gamma_j(t)}{\sqrt{C(t)\left(\frac{1}{t_{\mathrm{eff},i}(t)}+\frac{1}{t_{\mathrm{eff},j}(t)}\right)}}, 0, 1\right).\end{aligned}$$

For this lower bound to be at least q, the survival function must satisfy $\Phi(\cdot,0,1)\ge q$, which is equivalent to the argument being at most $\phi(q)$. By Cauchy–Schwarz,

$$\frac{\gamma_i(t)+\gamma_j(t)}{\sqrt{C(t)\left(\frac{1}{t_{\mathrm{eff},i}}+\frac{1}{t_{\mathrm{eff},j}}\right)}} \ \le\ \frac{\sqrt{2}\,(\sigma_0\sqrt{L_1(t)}+\sqrt{\rho}\,\varepsilon)}{\sqrt{C(t)}},$$

using $\gamma_i(t) \le (\sigma_0\sqrt{L_1(t)}+\sqrt{\rho}\,\varepsilon)/\sqrt{t_{\mathrm{eff},i}(t)}$ and $(a/\sqrt{x}+b/\sqrt{y})/\sqrt{1/x+1/y} \le \sqrt{2}\max(a,b)$ for the worst-case identical bias on both arms (more cleanly, $a/\sqrt{x}+b/\sqrt{y} \le \sqrt{2(a^2+b^2)}\sqrt{1/x+1/y}$, which we apply with $a=b=\sigma_0\sqrt{L_1}+\sqrt{\rho}\varepsilon$). Hence taking

$$C(t)\ :=\ \frac{2\big(\sigma_0\sqrt{L_1(t)}+\sqrt{\rho}\,\varepsilon\big)^2}{\phi^2(q)}$$

ensures the standardized shift is at most $\phi(q)$, so $\Pr\Big[\theta_i^m - \theta_j^m \ge \mu_i-\mu_j \mid G_0\Big] \ge q$ for every individual sample. (In the well-specified large-ρ regime $\sqrt{\rho}\,\varepsilon \le \sigma_0\sqrt{L_1}$ is automatic, in which case $C(t) = L_1(t)/\phi^2(q)$ up to absolute constants, recovering the choice in the main text.) Because the $M(\delta,q,t)$ samples are independent,

$$\Pr\big[\text{no sample index } m \text{ satisfies } \theta_i^m(t) - \theta_j^m(t) \ge \mu_i-\mu_j \mid G_0\big] \le (1-q)^{M(\delta,q,t)}.$$

[1] Keeping the original $6K^2$ denominator absorbs the $\pi^2/6$ factor without exceeding $\delta/3$. Relaxing to $\delta/(3Kt^2M)$ would have given $\sum_t \delta/(3t^2) = \delta\pi^2/18 \approx 0.548\delta$.

By the definition of $M(\delta, q, t)$,

$$(1-q)^{M(\delta,q,t)} \le \frac{\delta}{12K^2t^2}.$$

Taking a union bound over all ordered pairs (i, j) with $i \neq j$ and all $t \geq 1$ gives

$$\Pr(G_2^c \mid G_0) \le \sum_{t\ge1}\sum_{i\neq j} \frac{\delta}{12K^2t^2} \le \sum_{t\ge1} \frac{\delta}{6t^2} \le \frac{\delta}{3}.$$

Finally,

$$\Pr(G_0 \cap G_1 \cap G_2) \ge 1 - \Pr(G_0^c) - \Pr(G_1^c) - \Pr(G_2^c \mid G_0) \ge 1 - \delta.$$

This proves the claim. □

Proof of Theorem 2 (correctness). The correctness argument is brief once G_2 is established.

Proof: Assume the algorithm stops at time t and returns the empirical maximizer $\widehat{i}_t$. By the stopping rule, every Thompson sample agrees with the empirical maximizer, that is,

$$\widetilde{i}_t^m = \widehat{i}_t \qquad \text{for all } m = 1, \dots, M(\delta, q, t).$$

Suppose for contradiction that $\widehat{i}_t \neq a^\star$. Then $\mu_{a^\star} - \mu_{\widehat{i}_t} = \Delta_{\widehat{i}_t,c} > 0$. By event G_2, there exists some sample index m such that

$$\theta_{a^\star}^m(t) - \theta_{\widehat{i}_t}^m(t) \ge \mu_{a^\star} - \mu_{\widehat{i}_t} = \Delta_{\widehat{i}_t,c} > 0.$$

Therefore arm $a^\star$ has a strictly larger sampled value than $\widehat{i}_t$ in the m-th Thompson draw, which implies $\widetilde{i}_t^m \neq \widehat{i}_t$. This contradicts the stopping rule. Hence whenever the algorithm stops on G_2, the returned arm must equal $a^\star$. Since Lemma 1 shows $\Pr(G_0 \cap G_1 \cap G_2) \ge 1 - \delta$, the theorem follows. □

Proof of Lemma 3. This is the elimination step. The argument follows the classical TS-Explore proof, with $t_{\text{eff},i}(t)$ in place of the direct pull count. A sufficiently large effective count is enough to eliminate a suboptimal arm, even if that arm has not been sampled many times directly.

Proof: Fix a suboptimal arm $i \neq a^\star$ and suppose, toward contradiction, that arm i is still pulled at some time t even though

$$t_{\text{eff},i}(t) \ge \frac{186\, C(t)L_2(t)}{\Delta_{i,c}^2}. \tag{9}$$

Because the algorithm only ever pulls one of the two arms in the disagreement pair $\{\widehat{i}_t, \widetilde{i}_t\}$, the arm i must be either the empirical maximizer $\widehat{i}_t$ or the selected challenger $\widetilde{i}_t$.

We first consider the case $\widehat{i}_t = i$. Since i is selected for pulling, it is the less-sampled arm among $\{\widehat{i}_t, \widetilde{i}_t\}$, so

$$t_{\text{eff},\widetilde{i}_t}(t) \ge t_{\text{eff},i}(t) \ge \frac{186\, C(t)L_2(t)}{\Delta_{i,c}^2}.$$

On G_1, for every sample index m,

$$|\theta_i^m(t) - \widehat{\mu}_i(t)| \le \sqrt{\frac{2C(t)L_2(t)}{t_{\text{eff},i}(t)}} \le \frac{\Delta_{i,c}}{\sqrt{93}} < \frac{\Delta_{i,c}}{7},$$

and similarly,

$$|\theta^m_{\widetilde{i}_t}(t) - \widehat{\mu}_{\widetilde{i}_t}(t)| < \frac{\Delta_{i,c}}{7}.$$

Since $\widetilde{i}_t^{m^\star} = \widetilde{i}_t$ is the sampled challenger chosen to maximize the disagreement gap,

$$\theta^{m^\star}_{\widetilde{i}_t}(t) - \theta^{m^\star}_i(t) \geq 0.$$

Using the two bounds above, we obtain

$$\widehat{\mu}_{\widetilde{i}_t}(t) - \widehat{\mu}_i(t) \geq -\frac{2\Delta_{i,c}}{7}. \tag{10}$$

Now event G_2 provides a sample index m' such that

$$\theta^{m'}_{a^\star}(t) - \theta^{m'}_i(t) \geq \mu_{a^\star} - \mu_i = \Delta_{i,c}.$$

To convert this into a bound on the empirical means we need G_1 to control both $\theta^{m'}_{a^\star} - \widehat{\mu}_{a^\star}$ and $\theta^{m'}_i - \widehat{\mu}_i$ at deviation $\Delta_{i,c}/7$. The bound for arm i uses the assumption $t_{\mathrm{eff},i}(t) \geq 186\, C(t)L_2(t)/\Delta^2_{i,c}$ directly. The bound for $a^\star$ requires $t_{\mathrm{eff},a^\star}(t) \geq 186\, C(t)L_2(t)/\Delta^2_{i,c}$ as well. This holds at the relevant rounds by the following monotonicity. Order suboptimal arms by decreasing $\Delta_{j,c}$ (so smaller index means larger gap). For the smallest-index arm j_1, event G_2 guarantees a sample index m with $\theta^m_{a^\star} - \theta^m_{j_1} \geq \Delta_{j_1,c}$, so the maximum-disagreement challenger $\widetilde{i}_t$ has sampled value at least that of $a^\star$ on round t. Whenever $\widehat{i}_t = j_1$, the pull rule pulls one of $\{j_1, \widetilde{i}_t\}$, and $a^\star$ accumulates effective count whenever $\widetilde{i}_t = a^\star$ (and otherwise the same argument applied to $\widetilde{i}_t$ gives a lower bound on $t_{\mathrm{eff},\widetilde{i}_t}$ that propagates through the resistance metric). Iterating this argument over arms in order of increasing index $j_1, j_2, \ldots$ shows that by the time $t_{\mathrm{eff},i}(t)$ crosses its threshold, $t_{\mathrm{eff},a^\star}(t)$ has crossed the corresponding threshold for every arm j with $\Delta_{j,c} \geq \Delta_{i,c}$. Therefore G_1 yields

$$|\theta^{m'}_{a^\star}(t) - \widehat{\mu}_{a^\star}(t)| < \Delta_{i,c}/7 \quad \text{and} \quad |\theta^{m'}_i(t) - \widehat{\mu}_i(t)| < \Delta_{i,c}/7,$$

and we get

$$\widehat{\mu}_{a^\star}(t) - \widehat{\mu}_i(t) \geq \Delta_{i,c} - \frac{2\Delta_{i,c}}{7} = \frac{5\Delta_{i,c}}{7} > 0.$$

This contradicts the assumption that $i = \widehat{i}_t$ is the empirical maximizer.

We next consider the case $\widetilde{i}_t = i$. Then, since the pulled arm is chosen as the less-sampled arm in the pair $\{\widehat{i}_t, \widetilde{i}_t\}$, we again have

$$t_{\mathrm{eff},\widehat{i}_t}(t) \geq t_{\mathrm{eff},i}(t) \geq \frac{186\, C(t)L_2(t)}{\Delta^2_{i,c}}.$$

Therefore the same G_1 calculation yields

$$|\theta^m_i(t) - \widehat{\mu}_i(t)| < \frac{\Delta_{i,c}}{7}, \qquad |\theta^m_{\widehat{i}_t}(t) - \widehat{\mu}_{\widehat{i}_t}(t)| < \frac{\Delta_{i,c}}{7} \qquad \text{for all } m.$$

Because $\widetilde{i}_t = i$ is selected as the challenger with maximal sampled advantage over the empirical maximizer,

$$\theta^{m^\star}_i(t) - \theta^{m^\star}_{\widehat{i}_t}(t) \geq 0.$$

Hence

$$\widehat{\mu}_i(t) - \widehat{\mu}_{\widehat{i}_t}(t) \geq -\frac{2\Delta_{i,c}}{7}.$$

But G_2 again produces some sample index m' with

$$\theta_{a^\star}^{m'}(t) - \theta_i^{m'}(t) \geq \Delta_{i,c},$$

which implies

$$\widehat{\mu}_{a^\star}(t) - \widehat{\mu}_i(t) \geq \frac{5\Delta_{i,c}}{7}.$$

Combining the last two displays gives

$$\widehat{\mu}_{a^\star}(t) - \widehat{\mu}_{\widehat{i}_t}(t) \geq \frac{3\Delta_{i,c}}{7} > 0,$$

contradicting the fact that $\widehat{i}_t$ is the empirical maximizer.

Both cases lead to contradictions. Therefore a suboptimal arm cannot be pulled once its effective sample size crosses the threshold in (9). □

Proof of Theorem 5. We now convert the armwise elimination threshold into a bound on the total stopping time. The graph enters the complexity here through the lower bound on $t_{\text{eff},i}(t)$.

Proof: By Lemma 3, it is sufficient for successful elimination of every suboptimal arm i that

$$t_{\text{eff},i}(t) \geq \frac{196\, C(t) L_2(t)}{\Delta_{i,c}^2}. \tag{11}$$

We define

$$f_i(t) := \frac{196\, C(t) L_2(t)}{\Delta_{i,c}^2}.$$

Thus it is enough to ensure $t_{\text{eff},i}(t) \geq f_i(t)$ for all suboptimal arms.

We next apply Theorem 4, which gives

$$t_{\text{eff},i}(t) \geq t_i(t) + \frac{1}{2} \left\lfloor \min \left\{ \rho\, \mathcal{J}(i, G), \sum_{j \in C(i)} t_j(t) \right\} \right\rfloor.$$

Ignoring the floor for simplicity of constants, we obtain the softer but cleaner inequality

$$t_{\text{eff},i}(t) \gtrsim t_i(t) + \frac{1}{2} \min \left\{ \rho\, \mathcal{J}(i, G), \sum_{j \in C(i)} t_j(t) \right\}. \tag{12}$$

As in [47], this partitions the suboptimal arms into two classes.

Competitive arms. These are the arms for which the graph bonus alone is not enough to satisfy (11). Equivalently, the influence factor $\mathcal{J}(i, G)$ does not make

$$\rho\, \mathcal{J}(i, G) \gtrsim \frac{C(t) L_2(t)}{\Delta_{i,c}^2}$$

true at the relevant scale. Such arms must accumulate direct samples, and the threshold (11) therefore requires

$$t_i(t) \gtrsim \frac{C(t) L_2(t)}{\Delta_{i,c}^2}.$$

Summing over the competitive set $\mathcal{H} \setminus \{a^\star\}$ gives the direct-sampling contribution

$$\sum_{i \in \mathcal{H} \setminus \{a^\star\}} t_i(t) \lesssim C(t) L_2(t) \sum_{i \in \mathcal{H} \setminus \{a^\star\}} \frac{1}{\Delta_{i,c}^2}. \tag{13}$$

Non-competitive arms. For $i \in \mathcal{N}$, the graph bonus is large enough that once the total number of samples in the connected component of i exceeds a moderate threshold, the minimum term in (12) already exceeds $f_i(t)$. These arms are eliminated without requiring large per-arm direct pull counts. It suffices that the overall horizon satisfy

$$t \gtrsim C(t) L_2(t) \max_{i \in \mathcal{N}} \frac{1}{\Delta_{i,c}^2}.$$

This is the mechanism described in GRUB: non-competitive arms are eliminated quickly because their effective counts are inherited from their neighborhood.

Combining the two contributions, there is a numerical constant $c > 0$ such that the stopping time T is bounded by the smallest solution of

$$T \le c\, C(T) L_2(T) \left(\sum_{i \in \mathcal{H} \setminus \{a^\star\}} \frac{1}{\Delta_{i,c}^2} + \max_{i \in \mathcal{N}} \frac{1}{\Delta_{i,c}^2} \right). \tag{14}$$

The expression in parentheses is the graph hardness H_{graph} up to absolute constants. Since

$$C(T) = \frac{\log\big(12K^2T^2/\delta\big)}{\phi^2(q)} = O\left(\frac{\log(KT) + \log(1/\delta)}{\log(1/q)} \right),$$

and

$$L_2(T) = \log\left(\frac{12K^2T^2 M(\delta, q, T)}{\delta} \right) = O\big(\log(KT) + \log(1/\delta)\big),$$

substituting these into (14) yields

$$T = O\left(H_{\text{graph}} \left(\log \frac{1}{\delta} + \log(KT) \right) \frac{\log(1/\delta)}{\log(1/q)} + H_{\text{graph}} \frac{\log^2(KT)}{\log(1/q)} \right).$$

Finally, solving the self-referential logarithm in the standard way used in TS-Explore and LUCB analyses gives

$$T = O\left(H_{\text{graph}} \left(\log \frac{1}{\delta} + \log(K H_{\text{graph}}) \right) \frac{\log(1/\delta)}{\log(1/q)} + H_{\text{graph}} \frac{\log^2(K H_{\text{graph}})}{\log(1/q)} \right).$$

Setting $q = \delta$ reduces the first factor to a constant multiple of $\log(1/\delta)$ and yields the simplified bound

$$T = O\left(H_{\text{graph}} \log \frac{1}{\delta} + H_{\text{graph}} \log^2(K H_{\text{graph}}) \right).$$

This proves the theorem. □

D Proofs for the Optimal Regularization and Smoothness Asymptotics

We characterize the elimination threshold as a function of the regularization ρ, define the recommended scale $\rho_{\text{var}}(\varepsilon)$, prove Theorem 7, and establish the limiting behavior as $\varepsilon \to 0$ (Corollary 8). The analysis builds directly on the graph-smooth proofs in Appendix C. The good events G_0, G_1, G_2 and Lemma 1 hold for any choice of $\rho > 0$.

Notation. As before, we write $\sigma_0 = 2\sigma\sqrt{14}$, $L_1(t) = \log\big(12K^2t^2/\delta\big)$, $L_2(t) = \log\big(12K^2t^2M(\delta,q,t)/\delta\big)$, and $C(t) = 2(\sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon)^2/\phi^2(q)$ (the corrected variance scale derived in Appendix C. In the regime $\sqrt{\rho}\,\varepsilon \le \sigma_0\sqrt{L_1(t)}$ this reduces to $C(t) \asymp L_1(t)/\phi^2(q)$). The confidence width for the graph-smooth estimator is

$$\gamma_i(t) = \frac{1}{\sqrt{t_{\mathrm{eff},i}(t)}}\big(\sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon\big).$$

We abbreviate $\ell(t) := \sigma_0\sqrt{L_1(t)}$ and $\mathcal{J}_i := \mathcal{J}(i,G)$ for the influence factor of arm i.

Derivation of the non-competitive threshold. From the proof of Theorem 5 (Appendix C), arm i is eliminated once

$$t_{\mathrm{eff},i}(t) \;\ge\; \frac{186\,C(t)\,L_2(t)}{\Delta_{i,c}^2}. \tag{15}$$

By Theorem 4, the graph bonus alone yields $t_{\mathrm{eff},i}(t) \ge \rho\,\mathcal{J}_i/2$ (taking the minimum with the component pull count and ignoring floors). Thus, arm i is non-competitive: it is eliminated without needing its own direct pulls to dominate whenever

$$\frac{\rho\,\mathcal{J}_i}{2} \;\ge\; \frac{186\,C(t)\,L_2(t)}{\Delta_{i,c}^2}, \qquad \text{i.e.,} \qquad \Delta_{i,c}^2 \;\ge\; \frac{372\,C(t)\,L_2(t)}{\rho\,\mathcal{J}_i}. \tag{16}$$

To connect this with the threshold (4) in the main text, note that the variance scale $C(t) = L_1(t)/\phi^2(q)$ satisfies

$$C(t)\,L_2(t) \;\asymp\; L_1(t)\,L_2(t) \;\asymp\; (\ell(t))^2\,L_2(t)/\sigma_0^2.$$

Meanwhile, the G_2 event in Lemma 1 requires

$$\frac{\gamma_i(t) + \gamma_j(t)}{\sqrt{C(t)\big(\frac{1}{t_{\mathrm{eff},i}(t)} + \frac{1}{t_{\mathrm{eff},j}(t)}\big)}} \;\le\; \phi(q),$$

since the survival probability is $\Phi(\cdot,0,1) \ge q$ exactly when its argument is at most $\phi(q)$. Substituting the deterministic upper bound $\gamma_i(t) \le (\sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon)/\sqrt{t_{\mathrm{eff},i}(t)}$ and using $a/\sqrt{x} + a/\sqrt{y} \le \sqrt{2}\,a\sqrt{1/x + 1/y}$ with $a = \sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon$, the standardized shift is at most $\sqrt{2}\,(\sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon)/\sqrt{C(t)}$, which is at most $\phi(q)$ provided

$$C(t) \;\ge\; \frac{2\big(\sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon\big)^2}{\phi^2(q)}.$$

This is the corrected variance scale used throughout. Combining with (16), the non-competitive condition becomes

$$\Delta_{i,c} \;\gtrsim\; \frac{\sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon}{\sqrt{\rho\,\mathcal{J}_i/2}},$$

which is exactly (4).

Behavior of the non-competitive threshold in ρ.

Proof: [No finite interior $\rho^\star$.] Fix an arm $i \neq a^\star$ and consider the non-competitive threshold as a function of ρ:

$$f_i(\rho) := \frac{\ell(t) + \sqrt{\rho}\,\varepsilon}{\sqrt{\rho\,\mathcal{J}_i/2}} = \sqrt{\tfrac{2}{\mathcal{J}_i}}\left(\frac{\ell(t)}{\sqrt{\rho}} + \varepsilon\right).$$

Both terms in the parentheses are non-increasing in ρ (the first strictly), so $f_i(\rho)$ is strictly decreasing in ρ with infimum $\varepsilon\sqrt{2/\mathcal{J}_i}$ as $\rho \to \infty$. There is no interior optimum: increasing the regularization always weakens (i.e. improves) the non-competitive threshold.

This contrasts with a linear-in-ρ bias $\rho\varepsilon$, which would produce a balance point $\rho^\star = \ell/\varepsilon$. The correct $\sqrt{\rho}$ scaling, which follows from $V_t \succeq \rho L_G \Rightarrow V_t^{-1} \preceq \frac{1}{\rho} L_G^\dagger$ on range(L_G), changes the calculation: the variance term decays as $1/\sqrt{\rho}$ while the bias term is constant in ρ, giving a one-sided trade-off rather than an interior optimum. □

Proof: [Recommended ρ.] Although there is no finite optimum, the variance contribution $\ell/\sqrt{\rho}$ falls below the bias floor ε once $\rho \geq \rho_{\mathrm{var}}(\varepsilon) := \sigma_0^2 L_1(T)/\varepsilon^2$. Above this regularization scale, $f_i(\rho) \leq 2\varepsilon\sqrt{2/\mathcal{J}_i}$, and squaring gives the non-competitive condition

$$\Delta_{i,c}^2 \;\geq\; \frac{8\varepsilon^2}{\mathcal{J}(i,G)},$$

which is the characterization in Definition 6 with $c_0 = 8$. (Any finite c_0 obtained by absorbing the precise tail constants is equivalent up to the universal constants in Theorem 7.) □

Practical note on $L_1(T)$ in ρ_{var}. The right-hand side of (5) contains $L_1(T) = \log\big(12K^2T^2/\delta\big)$, which depends on the unknown stopping time T. The dependence is logarithmic, so a fixed pre-estimate based on a horizon proxy (e.g. a graph-free upper bound $T_0 := \sum_i \Delta_i^{-2}\log(1/\delta)$, or a doubling-trick on T) suffices. In our experiments we use a static estimate of L_1 at $T = 10^7$ (the per-run pull budget) so that ρ_{var} is a single numerical value. We treat the recommendation $\rho \geq \rho_{\mathrm{var}}(\varepsilon)$ as an order-of-magnitude prescription rather than an exact tuning rule. The empirical optima reported in Section 4 fall above this scale.

Good events under recommended regularization.

Proof: [Proof of the good-event lemma for $\rho \geq \rho_{\mathrm{var}}(\varepsilon)$.] The events G_0, G_1, G_2 from Lemma 1 hold for any $\rho > 0$ when $C(t)$ is taken as in this appendix. We verify each piece briefly.

Event G_0. The concentration inequality (Lemma A.2 of [47]) holds for any $\rho > 0$ with confidence width $\gamma_i(t) = (\ell(t) + \sqrt{\rho}\,\varepsilon)/\sqrt{t_{\mathrm{eff},i}(t)}$. At $\rho = \rho_{\mathrm{var}}(\varepsilon)$ the bias and variance pieces are equal in γ_i. At larger ρ the bias term dominates. The tail probability is $\delta/(2w_i(\boldsymbol{\pi}_t))$, and the union bound gives $\Pr(G_0^c) \leq \delta/3$ exactly as before.

Event G_1. Unchanged from Appendix C, as G_1 depends only on the Gaussian sampling step and not on ρ or ε. $\Pr(G_1^c) \leq \delta/3$.

Event G_2. With $C(t) = 2(\ell(t) + \sqrt{\rho}\,\varepsilon)^2/\phi^2(q)$ as defined above, the standardized shift satisfies

$$\frac{\gamma_i(t) + \gamma_j(t)}{\sqrt{C(t)(1/t_{\mathrm{eff},i} + 1/t_{\mathrm{eff},j})}} \leq \phi(q),$$

so each Thompson sample succeeds with probability $\geq q$, and the union bound gives $\Pr(G_2^c \mid G_0) \leq \delta/3$ as before.

Combining: $\Pr(G_0 \cap G_1 \cap G_2) \geq 1 - \delta$. □

Proof of Theorem 7 (sample complexity under recommended regularization).

Proof: The correctness argument (Theorem 2) is unchanged, since it uses only the G_2 event and the stopping rule.

For the sample complexity, the elimination threshold from Lemma 3 requires

$$t_{\mathrm{eff},i}(t) \;\geq\; \frac{186\, C(t)\, L_2(t)}{\Delta_{i,c}^2} \qquad \text{for all } i \neq a^\star.$$

We partition the arms using the competitive set from Definition 6.

Non-competitive arms ($i \in \mathcal{N}_\varepsilon$). By definition, $\Delta_{i,c}^2\, \mathcal{J}_i > c_0\, \varepsilon^2$. For any $\rho \geq \rho_{\mathrm{var}}(\varepsilon) = \sigma_0^2 L_1/\varepsilon^2$, the graph bonus is

$$\frac{\rho\, \mathcal{J}_i}{2} \;\geq\; \frac{\sigma_0^2 L_1\, \mathcal{J}_i}{2\varepsilon^2} \;>\; \frac{c_0(\sigma_0)^2 L_1}{2\Delta_{i,c}^2} \;\geq\; \frac{186\, C(t)\, L_2(t)}{\Delta_{i,c}^2}$$

for appropriate constants, using the corrected variance scale $C(t) \asymp (\sigma_0\sqrt{L_1} + \sqrt{\rho}\,\varepsilon)^2 \leq 4\sigma_0^2 L_1$ in the regime $\rho \leq \rho_{\mathrm{var}}(\varepsilon) \cdot O(1)$ where $\sqrt{\rho}\,\varepsilon \leq O(\sigma_0\sqrt{L_1})$ (if ρ is taken larger, the bias-floor characterization absorbs the additional log factors into c_0). Hence the elimination threshold is met by the graph bonus alone once the total horizon t is large enough that $\sum_{j \in C(i)} t_j \geq \rho\, \mathcal{J}_i$ (so that the minimum in Theorem 4 is saturated by the influence term). This requires $t \gtrsim C(t) L_2(t)/\Delta_{i,c}^2$. Across all non-competitive arms, the binding constraint is $t \gtrsim C(t) L_2(t) \cdot \max_{i \in \mathcal{N}_\varepsilon} \Delta_{i,c}^{-2}$.

Competitive arms ($i \in \mathcal{H}_\varepsilon \setminus \{a^\star\}$). The graph bonus does not suffice, so arm i must accumulate direct pulls. The threshold requires $t_i(t) \gtrsim C(t) L_2(t)/\Delta_{i,c}^2$. Summing: $\sum_{i \in \mathcal{H}_\varepsilon \setminus \{a^\star\}} t_i \lesssim C(t) L_2(t) \sum_{i \in \mathcal{H}_\varepsilon \setminus \{a^\star\}} \Delta_{i,c}^{-2}$.

Combining both contributions. The stopping time satisfies

$$T \;\leq\; c\, C(T)\, L_2(T) \left(\sum_{i \in \mathcal{H}_\varepsilon \setminus \{a^\star\}} \frac{1}{\Delta_{i,c}^2} + \max_{i \in \mathcal{N}_\varepsilon} \frac{1}{\Delta_{i,c}^2} \right) = c\, C(T)\, L_2(T) \cdot H_\varepsilon.$$

Since $C(T) = O\big((\log(KT) + \log(1/\delta))/\log(1/q)\big)$ and $L_2(T) = O\big(\log(KT) + \log(1/\delta)\big)$, substituting and solving the self-referential logarithm in the standard way yields

$$T = O\left(H_\varepsilon \left(\log \frac{1}{\delta} + \log(K H_\varepsilon) \right) \frac{\log(1/\delta)}{\log(1/q)} + H_\varepsilon \frac{\log^2(K H_\varepsilon)}{\log(1/q)} \right).$$

□

Proof of Corollary 8 (smoothness asymptotics).

Proof: Fix the arm set $[K]$, the gap vector $\{\Delta_{i,c}\}_{i \neq a^\star}$, and the graph G. Each suboptimal arm i has a critical smoothness level

$$\varepsilon_i^\star = \Delta_{i,c} \sqrt{\frac{\mathcal{J}(i, G)}{c_0}},$$

defined by the boundary of $\mathcal{H}_\varepsilon$ in Definition 6.

For any $\varepsilon < \varepsilon_i^\star$, arm i satisfies $\Delta_{i,c}^2\, \mathcal{J}_i > c_0 \varepsilon^2$, so $i \in \mathcal{N}_\varepsilon$. Conversely, for $\varepsilon \geq \varepsilon_i^\star$, $i \in \mathcal{H}_\varepsilon$.

Order the suboptimal arms so that $\varepsilon_{\pi(1)}^\star \geq \varepsilon_{\pi(2)}^\star \geq \cdots \geq \varepsilon_{\pi(K-1)}^\star$. As ε decreases past each $\varepsilon_{\pi(j)}^\star$, arm $\pi(j)$ moves from $\mathcal{H}_\varepsilon$ to $\mathcal{N}_\varepsilon$ and drops out of the sum in H_ε. Each such transition is a *phase transition* in the hardness.

Part (i). For any $\varepsilon < \varepsilon^\star_{\pi(K-1)} = \min_{i \neq a^\star} \varepsilon^\star_i$, every suboptimal arm has exited the competitive set, so $\mathcal{H}_\varepsilon = \{a^\star\}$ (or is empty of suboptimal arms). Hence $\mathcal{H}_\varepsilon \to \{a^\star\}$ as $\varepsilon \to 0$.

Part (ii). When $\mathcal{H}_\varepsilon \setminus \{a^\star\} = \varnothing$, the sum term vanishes and

$$H_\varepsilon = \max_{i \in \mathcal{N}_\varepsilon} \frac{1}{\Delta^2_{i,c}} = \max_{i \neq a^\star} \frac{1}{\Delta^2_{i,c}}.$$

Part (iii). Substituting $H_\varepsilon = \max_i \Delta_{i,c}^{-2}$ into Theorem 7 yields the stated bound. □

Remark on misspecification. If the algorithm is run with nominal smoothness budget ε while the true level is $\varepsilon_{\text{true}} \geq \varepsilon$, the bias term in γ_i becomes $\sqrt{\rho}\,\varepsilon_{\text{true}}$ rather than $\sqrt{\rho}\,\varepsilon$. For the survival argument of G_2 to remain valid, the Thompson variance scale $C(t)$ must be matched to an upper bound on the true smoothness. We treat the under-specified case in detail in Appendix G, showing that whenever such an upper bound $\overline{\varepsilon} \geq \varepsilon_{\text{true}}$ is available, the analysis above goes through with ε replaced by $\overline{\varepsilon}$.

E Proofs for the Graph-Feedback Model

We begin by restating the algorithm whose pseudocode was deferred from Section 3.3.

Algorithm 2 `TS-Explore-GF` (graph feedback)

Inputs: error constraint δ, threshold $q \in [\delta, 0.1]$, graph $G = ([n], E_G)$.
Initialize $N_i^{\text{fb}} \leftarrow 0$, $R_i^{\text{fb}} \leftarrow 0$ for all $i \in [n]$.
Pull a set S_0 of arms whose closed neighborhoods cover $[n]$ so that every arm is observed at least once; set $t \leftarrow |S_0|$ (the total number of initial pulls).
while true **do**
  $\widehat{\mu}_i(t) \leftarrow R_i^{\text{fb}}(t)/N_i^{\text{fb}}(t)$ for all i
  $\widehat{i}_t \leftarrow \arg\max_i \widehat{\mu}_i(t)$
  for $k = 1, \ldots, \lceil M(\delta, q, t) \rceil$ **do**
    for $i = 1, \ldots, n$ **do**
      Draw $\theta_i^k(t) \sim \mathcal{N}\big(\widehat{\mu}_i(t),\ C(\delta, q, t)/N_i^{\text{fb}}(t)\big)$
    end for
    $\widetilde{i}_t^k \leftarrow \arg\max_i \theta_i^k(t)$
    $\widehat{\Delta}_t^k \leftarrow \theta^k_{\widetilde{i}_t^k}(t) - \theta^k_{\widehat{i}_t}(t)$
  end for
  if $\widetilde{i}_t^k = \widehat{i}_t$ for all k **then**
    return $\widehat{i}_t$
  else
    $k^\star \leftarrow \arg\max_k \widehat{\Delta}_t^k$
    $\widetilde{i}_t \leftarrow \widetilde{i}_t^{k^\star}$
    Pull $\pi_{t+1} \leftarrow \arg\min_{a \in \{\widehat{i}_t, \widetilde{i}_t\}} N_a^{\text{fb}}(t)$, observe rewards from every $j \in N_G^+(\pi_{t+1})$, update $N_j^{\text{fb}}, R_j^{\text{fb}}$, and set $t \leftarrow t+1$.
  end if
end while

All proofs in this section follow those in Appendix C after replacing the graph-regularized effective sample size $t_{\text{eff},i}(t)$ with the feedback observation count $N_i^{\text{fb}}(t)$, and the Laplacian-regularized

estimator $\widehat{\mu}_i(t)$ with the empirical mean $R_i^{\mathrm{fb}}(t)/N_i^{\mathrm{fb}}(t)$. The simplification here is that each observation of arm i through the feedback graph is an independent draw from ν_i, so standard Hoeffding concentration applies directly with no Laplacian bias term.

We use the same shorthand as in Appendix C:

$$L_1(t) := \log\!\left(\frac{12n^2t^2}{\delta}\right), \qquad L_2(t) := \log\!\left(\frac{12n^2t^2M(\delta,q,t)}{\delta}\right), \qquad C(t) := C(\delta,q,t).$$

Definition of the good events. We work on the intersection of the following three events.

$$G_0^{\mathrm{fb}} := \Big\{\forall t \geq 1,\ \forall i \in [n],\ \big|\widehat{\mu}_i(t) - \mu_i\big| \leq \sqrt{\frac{L_1(t)}{2\,N_i^{\mathrm{fb}}(t)}}\Big\},$$

$$G_1^{\mathrm{fb}} := \Big\{\forall t \geq 1,\ \forall k \in [M(\delta,q,t)],\ \forall i \in [n],\ |\theta_i^k(t) - \widehat{\mu}_i(t)| \leq \sqrt{\frac{2C(t)L_2(t)}{N_i^{\mathrm{fb}}(t)}}\Big\},$$

$$G_2^{\mathrm{fb}} := \Big\{\forall t \geq 1,\ \forall i \neq j,\ \exists k \in [M(\delta,q,t)] \text{ s.t. } \theta_i^k(t) - \theta_j^k(t) \geq \mu_i - \mu_j\Big\}.$$

Event G_0^{fb} controls the empirical mean around the true mean (cf. G_0 in the graph-smooth model, which controls the Laplacian-regularized estimator). Because the estimator here is a simple sample average, no bias term from the graph smoothness appears. Event G_1^{fb} controls the Gaussian Thompson perturbations around the empirical means, and G_2^{fb} guarantees that among the $M(\delta,q,t)$ samples there is always at least one that separates any pair of arms by at least the true gap.

Proof of Lemma 9. We prove $\Pr\big(G_0^{\mathrm{fb}} \cap G_1^{\mathrm{fb}} \cap G_2^{\mathrm{fb}}\big) \geq 1-\delta$ by bounding the failure probability of each event separately.

Proof: *Bounding* $\Pr\big((G_0^{\mathrm{fb}})^c\big)$. Fix a time $t \geq 1$ and an arm $i \in [n]$. The feedback count $N_i^{\mathrm{fb}}(t)$ is itself a data-dependent random variable taking values in $\{0,1,\ldots,t\}$, so a fixed-sample concentration inequality cannot be applied directly. We therefore work with the deterministic peeling: for every $s \in \{1,\ldots,t\}$ Hoeffding's inequality gives

$$\Pr\left(\big|\bar{X}_i^{(s)} - \mu_i\big| > \sqrt{\frac{L_1(t)}{2s}}\right) \;\leq\; 2\exp(-L_1(t)) \;=\; \frac{\delta}{6n^2t^3},$$

where $\bar{X}_i^{(s)}$ is the average of the first s observations of arm i on the path realised by the algorithm, and where we use the strengthened deviation level

$$L_1(t) \;:=\; \log\!\left(\frac{12n^2t^3}{\delta}\right),$$

in which t^3 replaces the t^2 appearing in Appendix C. Taking a union bound over $s \in \{1,\ldots,t\}$,

$$\Pr\left(\big|\widehat{\mu}_i(t) - \mu_i\big| > \sqrt{\frac{L_1(t)}{2\,N_i^{\mathrm{fb}}(t)}}\right) \;\leq\; \sum_{s=1}^{t} \frac{\delta}{6n^2t^3} \;=\; \frac{\delta}{6n^2t^2},$$

which is the same per-(i,t) bound as in the deterministic case. Taking a further union bound over all n arms and all times $t \geq 1$,

$$\Pr\big((G_0^{\mathrm{fb}})^c\big) \leq \sum_{t\geq 1}\sum_{i=1}^{n} \frac{\delta}{6n^2t^2} \leq \sum_{t\geq 1} \frac{\delta}{6nt^2} \leq \sum_{t\geq 1} \frac{\delta}{6t^2} \leq \frac{\delta}{3}.$$

The same $\log(t^3)$ scaling is used wherever we invoke Hoeffding on a random-sample-size empirical mean in this appendix.

Bounding $\Pr\big((G_1^{\mathrm{fb}})^c\big)$. This step is identical to the G_1 argument in Appendix C, with $N_i^{\mathrm{fb}}(t)$ in place of $t_{\mathrm{eff},i}(t)$. Fix $t \geq 1$, $k \in [M(\delta, q, t)]$, and $i \in [n]$. Conditional on the history up to time t,

$$\theta_i^k(t) - \widehat{\mu}_i(t) \sim \mathcal{N}\left(0, \frac{C(t)}{N_i^{\mathrm{fb}}(t)}\right).$$

By standard Gaussian concentration,

$$\Pr\left[|\theta_i^k(t) - \widehat{\mu}_i(t)| > \sqrt{\frac{2C(t)L_2(t)}{N_i^{\mathrm{fb}}(t)}}\right] \leq 2e^{-L_2(t)} = \frac{\delta}{6n^2t^2M(\delta, q, t)}.$$

A union bound over arms, Thompson samples, and times gives

$$\Pr\big((G_1^{\mathrm{fb}})^c\big) \leq \sum_{t\geq 1}\sum_{k=1}^{M(\delta,q,t)}\sum_{i=1}^{n}\frac{\delta}{6n^2t^2M(\delta,q,t)} \leq \sum_{t\geq 1}\frac{\delta}{6nt^2} \leq \frac{\delta}{3}.$$

Bounding $\Pr\big((G_2^{\mathrm{fb}})^c \mid G_0^{\mathrm{fb}}\big)$. Fix $t \geq 1$ and two distinct arms i, j. Conditional on the empirical means,

$$\theta_i^k(t) - \theta_j^k(t) \sim \mathcal{N}\left(\widehat{\mu}_i(t) - \widehat{\mu}_j(t),\ C(t)\Big(\frac{1}{N_i^{\mathrm{fb}}(t)} + \frac{1}{N_j^{\mathrm{fb}}(t)}\Big)\right).$$

On G_0^{fb} we have $\widehat{\mu}_i(t) - \widehat{\mu}_j(t) \geq \mu_i - \mu_j - \sqrt{L_1(t)/(2N_i^{\mathrm{fb}}(t))} - \sqrt{L_1(t)/(2N_j^{\mathrm{fb}}(t))}$. Denote the sum of these confidence radii by $\gamma_i^{\mathrm{fb}}(t) + \gamma_j^{\mathrm{fb}}(t)$, where $\gamma_i^{\mathrm{fb}}(t) := \sqrt{L_1(t)/(2N_i^{\mathrm{fb}}(t))}$. Then, conditional on G_0^{fb},

$$\Pr\left[\theta_i^k(t) - \theta_j^k(t) \geq \mu_i - \mu_j \mid G_0^{\mathrm{fb}}\right] \geq \Phi\left(\frac{\gamma_i^{\mathrm{fb}}(t) + \gamma_j^{\mathrm{fb}}(t)}{\sqrt{C(t)\big(\frac{1}{N_i^{\mathrm{fb}}(t)} + \frac{1}{N_j^{\mathrm{fb}}(t)}\big)}},\ 0,\ 1\right).$$

By the same survival-function argument as in Appendix C, the right-hand side is at least q provided

$$\frac{\gamma_i^{\mathrm{fb}}(t) + \gamma_j^{\mathrm{fb}}(t)}{\sqrt{C(t)\big(\frac{1}{N_i^{\mathrm{fb}}(t)} + \frac{1}{N_j^{\mathrm{fb}}(t)}\big)}} \ \leq\ \phi(q),$$

since the standard normal survival function is strictly decreasing. Substituting $\gamma_i^{\mathrm{fb}}(t) = \sqrt{L_1(t)/(2N_i^{\mathrm{fb}}(t))}$ and using $\sqrt{a/x} + \sqrt{a/y} \leq \sqrt{2a}\sqrt{1/x + 1/y}$ with $a = L_1(t)/2$, the standardized shift is at most $\sqrt{L_1(t)/C(t)}$. Hence the choice

$$C(t) \ :=\ \frac{L_1(t)}{\phi^2(q)}$$

ensures the standardized shift is at most $\phi(q)$, so each Thompson draw succeeds with probability at least q. (Unlike the graph-smooth case, no bias term enters here, so this simpler $C(t)$ suffices.) Since the $M(\delta, q, t)$ samples are independent,

$$\Pr\left[\text{no } k \text{ satisfies } \theta_i^k(t) - \theta_j^k(t) \geq \mu_i - \mu_j \mid G_0^{\mathrm{fb}}\right] \leq (1-q)^{M(\delta,q,t)} \leq \frac{\delta}{12n^2t^2},$$

by the definition of $M(\delta, q, t)$. A union bound over all ordered pairs (i, j) with $i \neq j$ and all $t \geq 1$ gives

$$\Pr\big((G_2^{\mathrm{fb}})^c \mid G_0^{\mathrm{fb}}\big) \leq \sum_{t\geq 1} \sum_{i\neq j} \frac{\delta}{12n^2t^2} \leq \sum_{t\geq 1} \frac{\delta}{6t^2} \leq \frac{\delta}{3}.$$

Combining all three bounds,

$$\Pr\big(G_0^{\mathrm{fb}} \cap G_1^{\mathrm{fb}} \cap G_2^{\mathrm{fb}}\big) \geq 1 - \Pr\big((G_0^{\mathrm{fb}})^c\big) - \Pr\big((G_1^{\mathrm{fb}})^c\big) - \Pr\big((G_2^{\mathrm{fb}})^c \mid G_0^{\mathrm{fb}}\big) \geq 1 - \delta.$$

□

Proof of Theorem 10 (correctness).

Proof: Assume the algorithm stops at time t and returns the empirical maximizer $\widehat{i}_t$. By the stopping rule, $\widetilde{i}_t^k = \widehat{i}_t$ for all $k = 1, \ldots, M(\delta, q, t)$. Suppose for contradiction that $\widehat{i}_t \neq a^\star$. Then $\Delta_{\widehat{i}_t} = \mu_{a^\star} - \mu_{\widehat{i}_t} > 0$. By event G_2^{fb}, there exists some sample index k such that

$$\theta_{a^\star}^k(t) - \theta_{\widehat{i}_t}^k(t) \geq \mu_{a^\star} - \mu_{\widehat{i}_t} = \Delta_{\widehat{i}_t} > 0.$$

Therefore $\widetilde{i}_t^k \neq \widehat{i}_t$, contradicting the stopping rule. Hence, whenever the algorithm stops on $G_0^{\mathrm{fb}} \cap G_1^{\mathrm{fb}} \cap G_2^{\mathrm{fb}}$, the returned arm must be $a^\star$. Since $\Pr\big(G_0^{\mathrm{fb}} \cap G_1^{\mathrm{fb}} \cap G_2^{\mathrm{fb}}\big) \geq 1 - \delta$ by Lemma 9, the theorem follows. □

Proof of Lemma 11.

Proof: Fix a suboptimal arm $i \neq a^\star$ and suppose, toward contradiction, that arm i is still pulled at some time t even though

$$N_i^{\mathrm{fb}}(t) \geq \frac{186\, C(t)L_2(t)}{\Delta_i^2}. \tag{17}$$

Since the algorithm pulls one of the two arms in the disagreement pair $\{\widehat{i}_t, \widetilde{i}_t\}$ (specifically, the one with smaller current feedback count), arm i must appear in this pair. We consider both cases, mirroring the proof of Lemma 3 with $N_i^{\mathrm{fb}}(t)$ replacing $t_{\mathrm{eff},i}(t)$ and Δ_i replacing $\Delta_{i,c}$.

Case 1: $\widehat{i}_t = i$. Since i is the arm triggering the pull, $N_{\widetilde{i}_t}^{\mathrm{fb}}(t) \geq N_i^{\mathrm{fb}}(t) \geq 186\, C(t)L_2(t)/\Delta_i^2$. On G_1^{fb}, for every sample index k,

$$|\theta_i^k(t) - \widehat{\mu}_i(t)| \leq \sqrt{\frac{2C(t)L_2(t)}{N_i^{\mathrm{fb}}(t)}} \leq \frac{\Delta_i}{\sqrt{93}} < \frac{\Delta_i}{7},$$

and similarly $|\theta_{\widetilde{i}_t}^k(t) - \widehat{\mu}_{\widetilde{i}_t}(t)| < \Delta_i/7$. Since $\widetilde{i}_t$ is the challenger with maximal sampled advantage, $\theta_{\widetilde{i}_t}^{k^\star}(t) - \theta_i^{k^\star}(t) \geq 0$, which gives $\widehat{\mu}_{\widetilde{i}_t}(t) - \widehat{\mu}_i(t) \geq -2\Delta_i/7$. Now G_2^{fb} provides some index k' with $\theta_{a^\star}^{k'}(t) - \theta_i^{k'}(t) \geq \Delta_i$. Using the G_1^{fb} deviations,

$$\widehat{\mu}_{a^\star}(t) - \widehat{\mu}_i(t) \geq \Delta_i - \frac{2\Delta_i}{7} = \frac{5\Delta_i}{7} > 0,$$

contradicting the assumption that $i = \widehat{i}_t$ is the empirical maximizer.

Case 2: $\widetilde{i}_t = i$. Then $N_{\widehat{i}_t}^{\mathrm{fb}}(t) \geq N_i^{\mathrm{fb}}(t)$. The same G_1^{fb} calculation yields $|\theta_i^k(t) - \widehat{\mu}_i(t)| < \Delta_i/7$ and $|\theta_{\widehat{i}_t}^k(t) - \widehat{\mu}_{\widehat{i}_t}(t)| < \Delta_i/7$ for all k. Since $\widetilde{i}_t = i$ is the challenger, $\theta_i^{k^\star}(t) - \theta_{\widehat{i}_t}^{k^\star}(t) \geq 0$, so

$\widehat{\mu}_i(t) - \widehat{\mu}_{\widehat{i}_t}(t) \geq -2\Delta_i/7$. Event G_2^{fb} gives $\theta_{a^\star}^{k'}(t) - \theta_i^{k'}(t) \geq \Delta_i$, hence $\widehat{\mu}_{a^\star}(t) - \widehat{\mu}_i(t) \geq 5\Delta_i/7$. Combining,

$$\widehat{\mu}_{a^\star}(t) - \widehat{\mu}_{\widehat{i}_t}(t) \geq \frac{5\Delta_i}{7} - \frac{2\Delta_i}{7} = \frac{3\Delta_i}{7} > 0,$$

contradicting the fact that $\widehat{i}_t$ is the empirical maximizer.

Both cases lead to contradictions, so a suboptimal arm cannot be pulled once its feedback count crosses the threshold (17). □

Proof of Theorem 12.

Proof: By Lemma 11, successful elimination of every suboptimal arm i requires

$$N_i^{\mathrm{fb}}(t) \geq \frac{186\, C(t) L_2(t)}{\Delta_i^2} \qquad \text{for all } i \neq a^\star. \tag{18}$$

Define $f_i(t) := 186\, C(t) L_2(t)/\Delta_i^2$.

The structural difference from the graph-smooth model is the way per-arm counts are generated. In the graph-feedback model, a single action a simultaneously increments N_j^{fb} for every $j \in N_G^+(a)$. Let τ_a denote the number of times action a is pulled up to time t. Then $N_i^{\mathrm{fb}}(t) = \sum_{a:\, i \in N_G^+(a)} \tau_a$. The sufficient condition (18) can therefore be written as

$$\sum_{a:\, i \in N_G^+(a)} \tau_a \geq f_i(t) \qquad \text{for all } i \neq a^\star.$$

We compare the realized total T to the value of the covering LP

$$H_{\mathrm{GF}}(\boldsymbol{f}) \;:=\; \min_{\boldsymbol{\tau} \in \mathbb{R}_+^n} \left\{ \sum_{a=1}^n \tau_a : \sum_{a:\, i \in N_G^+(a)} \tau_a \geq f_i(t) \text{ for all } i \neq a^\star \right\},$$

which corresponds to the optimal allocation of pulls if any arm in $[n]$ were available for pulling at every round. Because $f_i(t) = 186\, C(t) L_2(t)/\Delta_i^2$, the value of this LP equals $186\, C(T) L_2(T) \cdot H_{\mathrm{GF}}$ with H_{GF} as in (7).

The pull rule of Algorithm 2 restricts π_t to $\{\widehat{i}_t, \widetilde{i}_t\}$, so the realized allocation is in general suboptimal relative to the unrestricted LP. The relationship between the restricted and unrestricted LP values is the standard *set-cover integrality gap*: any feasible cover for the unrestricted LP can be converted into a cover by elements of $[n]$ by taking each element as one of its own neighbors, but doing so may increase the cost by an $O(\log n)$ factor in the worst case (cf. Wolsey's $\ln(\Delta + 1)$ bound for greedy and restricted covering, where Δ is the maximum degree of any arm in any candidate's closed neighborhood). Thus

$$T \leq 186\, C(T) L_2(T) \cdot H_{\mathrm{GF}} \cdot O(\log n).$$

Since

$$C(T) = \frac{L_1(T)}{\phi^2(q)} = O\!\left(\frac{\log(nT) + \log(1/\delta)}{\log(1/q)}\right), \qquad L_2(T) = O\big(\log(nT) + \log(1/\delta)\big),$$

substituting into the bound above yields

$$T = O\!\left(H_{\mathrm{GF}} \log n \left(\log \frac{1}{\delta} + \log(nT)\right) \frac{\log(1/\delta)}{\log(1/q)} + H_{\mathrm{GF}} \log n \, \frac{\log^2(nT)}{\log(1/q)}\right).$$

Solving the self-referential logarithm in the standard way (as in the TS-Explore and LUCB analyses) gives

$$T = O\left(H_{\mathrm{GF}} \log n \left(\log \frac{1}{\delta} + \log(n H_{\mathrm{GF}}) \right) \frac{\log(1/\delta)}{\log(1/q)} + H_{\mathrm{GF}} \log n \, \frac{\log^2(n H_{\mathrm{GF}})}{\log(1/q)} \right).$$

Setting $q = \delta$ reduces this to $O\big(H_{\mathrm{GF}} \log n \log(1/\delta) + H_{\mathrm{GF}} \log n \log^2(n H_{\mathrm{GF}})\big)$.

Tightness of the $\log n$ *factor.* The factor $\log n$ is the worst-case integrality gap of restricting the cover to candidate vertices in a disagreement pair. On graphs where the chosen vertices already cover the full elimination requirement (e.g. when each disagreement candidate is itself a low-degree arm in a homogeneous expander), the factor is $O(1)$. We do not attempt to remove it here as replacing the pull rule with an LP-rounded action choice would be required. □

Proof of Corollary 13.

Proof: We separately prove the three inequalities in (8) and the duality statement.

Min-degree upper bound (right-most inequality). Let $c := \Delta_{\max}^{-2}/(1 + d_{\min}(G))$ and set $\tau_a = c$ for all $a \in [n]$. For each $i \neq a^\star$,

$$\sum_{a:\, i \in N_G^+(a)} \tau_a = c\,|N_G^+(i)| = c\,(1 + \deg_G(i)) \geq c\,(1 + d_{\min}(G)) = \Delta_{\max}^{-2} \geq \Delta_i^{-2},$$

so $\boldsymbol{\tau}$ is feasible for the LP in (7). The objective value is $\sum_a \tau_a = nc = n\Delta_{\max}^{-2}/(1 + d_{\min}(G))$.

Clique-cover upper bound (middle inequality). Let $V_G = C_1 \sqcup \cdots \sqcup C_{\bar{\chi}(G)}$ be a partition of V_G into cliques and pick a representative $a_j \in C_j$ for each j. Set $\tau_{a_j} = \Delta_{\max}^{-2}$ and $\tau_a = 0$ for $a \notin \{a_1, \ldots, a_{\bar{\chi}(G)}\}$. For each suboptimal $i \in C_j$, since C_j is a clique we have $a_j \in N_G^+(i)$, hence

$$\sum_{a:\, i \in N_G^+(a)} \tau_a \geq \tau_{a_j} = \Delta_{\max}^{-2} \geq \Delta_i^{-2}.$$

The objective value of this feasible $\boldsymbol{\tau}$ is $\bar{\chi}(G) \cdot \Delta_{\max}^{-2}$.

Packing lower bound (left-most inequality). Let $S^\circ \subseteq [n]$ achieve the 2-packing number, $|S^\circ| = \rho(G)$, and set $S := S^\circ \setminus \{a^\star\}$, so $|S| \geq \rho(G) - 1$ and the closed neighborhoods $\{N_G^+(i) : i \in S\}$ remain pairwise disjoint. For any feasible $\boldsymbol{\tau}$ for the LP (7),

$$\sum_{a=1}^{n} \tau_a \;\geq\; \sum_{i \in S} \sum_{a \in N_G^+(i)} \tau_a \;\geq\; \sum_{i \in S} \Delta_i^{-2} \;\geq\; |S| \cdot \Delta_{\min}^{-2} \;\geq\; (\rho(G) - 1)\, \Delta_{\min}^{-2},$$

where the first inequality uses pairwise disjointness of the closed neighborhoods (each τ_a contributes to at most one term in the inner sum) and the second uses feasibility of $\boldsymbol{\tau}$. When $a^\star \notin S^\circ$ the bound $\rho(G)\,\Delta_{\min}^{-2}$ in (8) is recovered. The -1 accounts for the worst case in which $a^\star$ lies in every maximum 2-packing and removing it shrinks $|S|$ by one.

LP duality in the equal-gap case. When $\Delta_i \equiv \Delta$ for all $i \neq a^\star$, the LP (7) (after factoring out Δ^{-2}) reads

$$\Delta^{-2} \cdot \min_{\boldsymbol{\tau} \geq 0} \Big\{ \mathbf{1}^\top \boldsymbol{\tau} \;:\; B\boldsymbol{\tau} \geq \mathbf{1} \Big\}, \qquad B_{ia} = \mathbf{1}\{i \in N_G^+(a)\} \;\; (i \neq a^\star,\; a \in [n]),$$

which is the LP relaxation of the dominating-set problem on G restricted to suboptimal arms. By LP duality, its optimal value equals

$$\max_{\boldsymbol{y} \geq 0} \Big\{ \mathbf{1}^\top \boldsymbol{y} \;:\; B^\top \boldsymbol{y} \leq \mathbf{1} \Big\} = \max_{\boldsymbol{y} \geq 0} \Big\{ \sum_{i \neq a^\star} y_i \;:\; \sum_{i \in N_G^+(a)} y_i \leq 1 \;\forall a \in [n] \Big\},$$

which is exactly the fractional 2-packing problem on G. Hence $H_{\mathrm{GF}} = \Delta^{-2}\rho^*(G)$. □

Worked examples for Corollary 13. We verify the values quoted in the main text under uniform gaps $\Delta_i \equiv \Delta$ for $i \neq a^\star$. Table 1 summarizes the resulting hardness on canonical graph families.

Table 1: Closed-form H_{GF} on canonical graphs under uniform gap Δ, with the matching graph parameters from Corollary 13.

Graph G	H_{GF}	$\rho(G)$	$\bar{\chi}(G)$
Empty graph	$(n-1)\,\Delta^{-2}$	$n-1$	$n-1$
Complete graph K_n	Δ^{-2}	1	1
Star $K_{1,n-1}$	Δ^{-2}	1	1
d-regular (clique union)	$n\,\Delta^{-2}/(d+1)$	$n/(d+1)$	$n/(d+1)$
d-regular (expander)	$n\,\Delta^{-2}/(d+1)$	$\ll n/(d+1)$	$n/(d+1)$

Empty graph. The constraints in (7) decouple into $\tau_i \geq \Delta^{-2}$ for each $i \neq a^\star$, so $H_{\mathrm{GF}} = (n-1)\Delta^{-2}$. Both bounds in (8) are tight: each closed neighborhood is a singleton, hence $\rho(G) = n-1$ and $\bar{\chi}(G) = n-1$.

Complete graph K_n. Setting $\tau_1 = \Delta^{-2}$ and $\tau_a = 0$ for $a \geq 2$ is feasible because $1 \in N_G^+(i)$ for every $i \in [n]$. This gives $H_{\mathrm{GF}} \leq \Delta^{-2}$. Conversely, $H_{\mathrm{GF}} \geq \Delta^{-2}$ from the constraint for any single suboptimal arm. Hence $H_{\mathrm{GF}} = \Delta^{-2}$, with $\bar{\chi}(G) = \rho(G) = 1$.

Star $K_{1,n-1}$. Let c be the center and $\ell_1, \ldots, \ell_{n-1}$ the leaves. Setting $\tau_c = \Delta^{-2}$ and $\tau_{\ell_k} = 0$ for all k is feasible: for the center constraint $\sum_{a \in N_G^+(c)} \tau_a = \Delta^{-2}$, and for each leaf ℓ_k, $c \in N_G^+(\ell_k)$ so $\sum_{a \in N_G^+(\ell_k)} \tau_a \geq \tau_c = \Delta^{-2}$. Hence $H_{\mathrm{GF}} \leq \Delta^{-2}$, with equality from the trivial single-arm lower bound.

d-regular graph. The uniform allocation $\tau_a \equiv \Delta^{-2}/(d+1)$ is feasible since each closed neighborhood has exactly $d+1$ vertices, giving $\sum_{a \in N_G^+(i)} \tau_a = (d+1) \cdot \Delta^{-2}/(d+1) = \Delta^{-2}$ for each i. The objective value is $n\Delta^{-2}/(d+1)$. On a disjoint union of K_{d+1}'s the clique-cover number is $\bar{\chi}(G) = n/(d+1)$, so the upper bound is matched by the trivial cover-by-cliques lower bound. On expander-like d-regular graphs the 2-packing number $\rho(G)$ can be much smaller than $n/(d+1)$, leaving a gap between the two bounds in (8).

F Proofs for the Normalized-Laplacian Kernel Model

We begin by restating the algorithm whose pseudocode was deferred from Section 3.4.

Algorithm 3 `Kernel-TS-Explore`

Inputs: graph G, PSD kernel K_G, regularization parameter $\rho > 0$, error constraint δ, threshold $q \in [\delta, 0.1]$.
Pull each arm $i \in [n]$ once and set $t \leftarrow n$ (so the index s in V_t ranges over the n initial pulls).
Set $V_t \leftarrow \sum_{s=1}^{n} e_{\pi_s} e_{\pi_s}^\top + \rho K_G, \quad \mathbf{x}_t \leftarrow \sum_{s=1}^{t} e_{\pi_s} r_{s,\pi_s}, \quad \widehat{\boldsymbol{\mu}}(t) \leftarrow V_t^{-1}\mathbf{x}_t.$
while true **do**
  $\widehat{i}_t \leftarrow \arg\max_{i\in[n]} \widehat{\mu}_i(t)$
  for $m = 1, \dots, \lceil M(\delta, q, t)\rceil$ **do**
    for $i = 1, \dots, n$ **do**
      Compute $t_{\mathrm{eff},i}(t) = \big[(V_t^{-1})_{ii}\big]^{-1}$
      Sample $\theta_i^m(t) \sim \mathcal{N}\big(\widehat{\mu}_i(t),\ C(\delta, q, t)/t_{\mathrm{eff},i}(t)\big)$
    end for
    $\widetilde{i}_t^m \leftarrow \arg\max_{i\in[n]} \theta_i^m(t)$
    $\widehat{\Delta}_t^m \leftarrow \theta_{\widetilde{i}_t^m}^m(t) - \theta_{\widehat{i}_t}^m(t)$
  end for
  if $\widetilde{i}_t^m = \widehat{i}_t$ for all $m = 1, \dots, \lceil M(\delta, q, t)\rceil$ **then**
    return $\widehat{i}_t$
  else
    $m^\star \leftarrow \arg\max_m \widehat{\Delta}_t^m$
    $\widetilde{i}_t \leftarrow \widetilde{i}_t^{m^\star}$
    Pull $\pi_{t+1} \leftarrow \arg\min_{i\in\{\widehat{i}_t, \widetilde{i}_t\}} t_{\mathrm{eff},i}(t)$
    Observe $r_{t+1,\pi_{t+1}}$, update $V_{t+1} \leftarrow V_t + e_{\pi_{t+1}} e_{\pi_{t+1}}^\top$, $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + r_{t+1,\pi_{t+1}} e_{\pi_{t+1}}$, $\widehat{\boldsymbol{\mu}}(t+1) \leftarrow V_{t+1}^{-1}\mathbf{x}_{t+1}$, and set $t \leftarrow t+1$.
  end if
end while

Setup. The estimator in this model is

$$\widehat{\boldsymbol{\mu}}(t) = V_t^{-1} \sum_{s=1}^{t} e_{\pi_s} r_{s,\pi_s}, \qquad V_t = \sum_{s=1}^{t} e_{\pi_s} e_{\pi_s}^\top + \rho K_G,$$

with kernel-effective sample sizes $t_{\mathrm{eff},i}^{(\mathrm{norm})}(t) := [(V_t)^{-1}]_{ii}^{-1}$. The smoothness assumption is $\langle \boldsymbol{\mu}, K_G \boldsymbol{\mu}\rangle \le \varepsilon^2$. We define the confidence width

$$\gamma_i^{(\mathrm{norm})}(t) := \sqrt{\frac{1}{t_{\mathrm{eff},i}^{(\mathrm{norm})}(t)}} \left(2\sigma \sqrt{14 \log\!\left(\frac{2w_i(\boldsymbol{\pi}_t)}{\delta}\right)} + \sqrt{\rho}\, \|\boldsymbol{\mu}\|_{K_G} \right),$$

with the deterministic upper bound $\|\boldsymbol{\mu}\|_{K_G} \le \varepsilon$.

Definition of the good events.

$$G_0^{(\mathrm{norm})} := \Big\{ \forall t \ge 1,\ \forall i \in [K],\ |\widehat{\mu}_i(t) - \mu_i| \le \gamma_i^{(\mathrm{norm})}(t) \Big\},$$

$$G_1^{(\mathrm{norm})} := \Big\{ \forall t \ge 1,\ \forall m \in [M(\delta, q, t)],\ \forall i \in [K],\ |\theta_i^m(t) - \widehat{\mu}_i(t)| \le \sqrt{\frac{2C(t)L_2(t)}{t_{\mathrm{eff},i}^{(\mathrm{norm})}(t)}} \Big\},$$

$$G_2^{(\mathrm{norm})} := \Big\{ \forall t \ge 1,\ \forall i \ne j,\ \exists m \in [M(\delta, q, t)] \text{ s.t. } \theta_i^m(t) - \theta_j^m(t) \ge \mu_i - \mu_j \Big\}.$$

Proof of the good-event lemma (analogue of Lemma 1). We show $\Pr\big(G_0^{(\mathrm{norm})} \cap G_1^{(\mathrm{norm})} \cap G_2^{(\mathrm{norm})}\big) \geq 1-\delta$.

Proof: *Bounding* $\Pr\Big((G_0^{(\mathrm{norm})})^c\Big)$. The estimator $\widehat{\boldsymbol{\mu}}(t) = V_t^{-1}\sum_{s=1}^{t} e_{\pi_s} r_{s,\pi_s}$ is a ridge-type solution with PSD penalty ρK_G. Since $K_G \succeq 0$, the matrix $V_t = \sum_{s=1}^{t} e_{\pi_s} e_{\pi_s}^\top + \rho K_G$ is positive definite once at least one arm per connected component has been sampled. Writing the estimation error for arm i as

$$\widehat{\mu}_i(t) - \mu_i = \big\langle e_i,\, V_t^{-1}\sum_{s=1}^{t} e_{\pi_s}(r_{s,\pi_s} - \mu_{\pi_s})\big\rangle - \big\langle e_i,\, \rho V_t^{-1} K_G \boldsymbol{\mu}\big\rangle,$$

the first (noise) term is a martingale in the filtration, and the second (bias) term is controlled deterministically. For the noise term, we apply the self-normalized martingale inequality (the analogue of Lemma A.2 of [47]): for any fixed t and i,

$$\Pr\left(\left|\big\langle e_i,\, V_t^{-1}\sum_{s=1}^{t} e_{\pi_s}(r_{s,\pi_s} - \mu_{\pi_s})\big\rangle\right| > \sqrt{[V_t^{-1}]_{ii}}\cdot 2\sigma\sqrt{14\log\left(\frac{2w_i(\boldsymbol{\pi}_t)}{\delta}\right)}\right) \leq \frac{\delta}{2w_i(\boldsymbol{\pi}_t)}.$$

This inequality depends only on the PSD structure of V_t and the sub-Gaussian property of the noise, not on whether the penalty is L_G or K_G. For the bias term, the Cauchy–Schwarz inequality gives

$$\big|\langle e_i, \rho V_t^{-1} K_G \boldsymbol{\mu}\rangle\big| \leq \rho\sqrt{[V_t^{-1}]_{ii}}\,\sqrt{\langle K_G\boldsymbol{\mu}, V_t^{-1} K_G \boldsymbol{\mu}\rangle} \leq \sqrt{\rho}\,\sqrt{[V_t^{-1}]_{ii}}\,\|\boldsymbol{\mu}\|_{K_G},$$

where the last step uses $V_t \succeq \rho K_G$, hence $V_t^{-1} \preceq \frac{1}{\rho}K_G^{\dagger}$ on the range of K_G, so that $\langle K_G\boldsymbol{\mu}, V_t^{-1}K_G\boldsymbol{\mu}\rangle \leq \frac{1}{\rho}\langle \boldsymbol{\mu}, K_G\boldsymbol{\mu}\rangle = \|\boldsymbol{\mu}\|_{K_G}^2/\rho$. Combining both terms,

$$|\widehat{\mu}_i(t) - \mu_i| \leq \gamma_i^{(\mathrm{norm})}(t)$$

holds except with probability $\delta/(2w_i(\boldsymbol{\pi}_t))$. Choosing $w_i(\boldsymbol{\pi}_t) = a_0 n(t_{\mathrm{eff},i}^{(\mathrm{norm})}(t))^2$ for an appropriate constant $a_0 > 0$ and applying a union bound over arms and times exactly as in the proof for G_0 in Appendix C yields $\Pr\Big((G_0^{(\mathrm{norm})})^c\Big) \leq \delta/3$.

Bounding $\Pr\Big((G_1^{(\mathrm{norm})})^c\Big)$. This step is identical to the G_1 proof in Appendix C. Fix t, m, and i. Conditional on the history,

$$\theta_i^m(t) - \widehat{\mu}_i(t) \sim \mathcal{N}\left(0, \frac{C(t)}{t_{\mathrm{eff},i}^{(\mathrm{norm})}(t)}\right).$$

By Gaussian concentration,

$$\Pr\left[|\theta_i^m(t) - \widehat{\mu}_i(t)| > \sqrt{\frac{2C(t)L_2(t)}{t_{\mathrm{eff},i}^{(\mathrm{norm})}(t)}}\right] \leq 2e^{-L_2(t)} = \frac{\delta}{6K^2t^2M(\delta,q,t)}.$$

A union bound over arms, samples, and times gives $\Pr\Big((G_1^{(\mathrm{norm})})^c\Big) \leq \delta/3$.

Bounding $\Pr\Big((G_2^{(\mathrm{norm})})^c \mid G_0^{(\mathrm{norm})}\Big)$. Fix t and two arms i, j. Conditional on the empirical means,

$$\theta_i^m(t) - \theta_j^m(t) \sim \mathcal{N}\left(\widehat{\mu}_i(t) - \widehat{\mu}_j(t),\; C(t)\Big(\frac{1}{t_{\mathrm{eff},i}^{(\mathrm{norm})}(t)} + \frac{1}{t_{\mathrm{eff},j}^{(\mathrm{norm})}(t)}\Big)\right).$$

On $G_0^{(\mathrm{norm})}$, $\widehat{\mu}_i(t) - \widehat{\mu}_j(t) \geq \mu_i - \mu_j - \gamma_i^{(\mathrm{norm})}(t) - \gamma_j^{(\mathrm{norm})}(t)$. The same Gaussian survival-function argument as in the G_2 step of Appendix C shows that each Thompson draw satisfies the pairwise comparison with probability at least q. Independence over the $M(\delta, q, t)$ draws, followed by a union bound over pairs and times, gives $\Pr\Big((G_2^{(\mathrm{norm})})^c \mid G_0^{(\mathrm{norm})}\Big) \leq \delta/3$.

Combining all three bounds, $\Pr\Big(G_0^{(\mathrm{norm})} \cap G_1^{(\mathrm{norm})} \cap G_2^{(\mathrm{norm})}\Big) \geq 1 - \delta$. □

Correctness (analogue of Theorem 2).

Proof: The argument is identical to the proof of Theorem 2. Assume the algorithm stops at time t and returns $\widehat{i}_t$. If $\widehat{i}_t \neq a^\star$, then $G_2^{(\mathrm{norm})}$ provides a sample index m with $\theta_{a^\star}^m(t) - \theta_{\widehat{i}_t}^m(t) \geq \mu_{a^\star} - \mu_{\widehat{i}_t} > 0$, so $\widetilde{i}_t^m \neq \widehat{i}_t$, contradicting the stopping rule. Hence the returned arm is $a^\star$ on the good event, which holds with probability at least $1 - \delta$. □

Elimination threshold (analogue of Lemma 3).

Proof: Fix a suboptimal arm $i \neq a^\star$ and suppose i is still pulled at time t despite $t_{\mathrm{eff},i}^{(\mathrm{norm})}(t) \geq 186\, C(t) L_2(t)/\Delta_{i,c}^2$. Both cases ($\widehat{i}_t = i$ and $\widetilde{i}_t = i$) proceed exactly as in the proof of Lemma 3, using $t_{\mathrm{eff},i}^{(\mathrm{norm})}(t)$ in place of $t_{\mathrm{eff},i}(t)$. On $G_1^{(\mathrm{norm})}$, the Thompson deviations satisfy $|\theta_i^m(t) - \widehat{\mu}_i(t)| < \Delta_{i,c}/7$, and $G_2^{(\mathrm{norm})}$ provides the requisite separation. Both cases lead to contradictions with the definition of $\widehat{i}_t$, so arm i cannot be pulled once its kernel-effective count exceeds the threshold. □

F.1 Spectral characterization of kernel-effective counts

Before turning to the sample-complexity argument we record the spectral content of $t_{\mathrm{eff},i}$, which is the basis for Proposition 15 in the main text. The variational form converts the matrix-inversion question into a constrained quadratic minimization that makes both the counting matrix and the kernel explicit.

Lemma 16 (Variational form of t_{eff}) *For any sampling counts $\{t_j\}_{j=1}^n$ with $V_T \succ 0$ and any arm $i \in [n]$,*

$$t_{\mathrm{eff},i}(T) = \min_{\substack{x \in \mathbb{R}^n \\ x_i = 1}} \left\{ \sum_{j=1}^n t_j\, x_j^2 + \rho\, \langle x, K_G\, x \rangle \right\}.$$

Proof: The right-hand side equals $\min_{x:\, e_i^\top x = 1} x^\top V_T x$. The Lagrangian $L(x, \nu) = x^\top V_T x - 2\nu(e_i^\top x - 1)$ has stationarity condition $V_T x = \nu e_i$, so $x^\star = \nu V_T^{-1} e_i$. Imposing $e_i^\top x^\star = 1$ gives $\nu = 1/[V_T^{-1}]_{ii}$. Substituting,

$$\min = (x^\star)^\top V_T x^\star = \nu^2\, e_i^\top V_T^{-1} e_i = \nu^2\, [V_T^{-1}]_{ii} = \frac{1}{[V_T^{-1}]_{ii}} = t_{\mathrm{eff},i}(T).$$

□

Proposition 17 (Spectral lower bound) *Let K_G be PSD with $\ker(K_G) = \operatorname{span}(v_0)$ for some unit vector $v_0 \in \mathbb{R}^n$. Let $\lambda_2(K_G) > 0$ denote the smallest positive eigenvalue of K_G. For any $\{t_j\}$ with $T = \sum_j t_j > 0$ and every $i \in [n]$,*

$$\frac{1}{t_{\mathrm{eff},i}(T)} \;\le\; \frac{1}{t_i + \rho\lambda_2(K_G)} \;+\; \frac{\rho\lambda_2(K_G)\,(v_0)_i^2}{\big(t_i + \rho\lambda_2(K_G)\big)^2\,(1-\varepsilon_T)},$$

where $\varepsilon_T := \rho\lambda_2(K_G)\sum_{j=1}^n \frac{(v_0)_j^2}{t_j+\rho\lambda_2(K_G)} \in [0,1)$. For the combinatorial Laplacian L_G, $v_0 = \mathbf{1}/\sqrt{n}$, recovering $(v_0)_i^2 = 1/n$ uniformly. For the normalized Laplacian $\mathcal{L}_G = I - D_G^{-1/2} A_G D_G^{-1/2}$, $\ker(\mathcal{L}_G) = \operatorname{span}(D_G^{1/2}\mathbf{1})$, so $v_0 = D_G^{1/2}\mathbf{1}/\|D_G^{1/2}\mathbf{1}\|$ and $(v_0)_i^2 = \deg_G(i)/(2|E_G|)$ varies with the degree.

Proof: Since K_G is PSD with kernel $\operatorname{span}(v_0)$, every $u \perp v_0$ satisfies $\langle u, K_G u\rangle \ge \lambda_2(K_G)\|u\|_2^2$, i.e., $K_G \succeq \lambda_2(K_G)\,P^\perp$ where $P^\perp = I - v_0 v_0^\top$ is the projection onto $v_0^\perp$. Therefore

$$V_T \;=\; N_T + \rho K_G \;\succeq\; N_T + \rho\lambda_2(K_G)\,P^\perp \;=\; D - \rho\lambda_2(K_G)\,v_0 v_0^\top \;=:\; M,$$

where $D = \operatorname{diag}\big(t_j + \rho\lambda_2(K_G)\big)_{j=1}^n$. By operator-monotonicity of the inverse, $[V_T^{-1}]_{ii} \le [M^{-1}]_{ii}$. Apply the Sherman–Morrison formula to $M = D - \alpha\, v_0 v_0^\top$ with $\alpha = \rho\lambda_2(K_G)$:

$$M^{-1} = D^{-1} + \frac{\alpha\, D^{-1} v_0 v_0^\top D^{-1}}{1 - \alpha\, v_0^\top D^{-1} v_0}.$$

Reading off the (i,i)-entry,

$$[M^{-1}]_{ii} = \frac{1}{t_i + \rho\lambda_2(K_G)} + \frac{\alpha\,(v_0)_i^2/(t_i+\rho\lambda_2(K_G))^2}{1 - \alpha\sum_j (v_0)_j^2/(t_j+\rho\lambda_2(K_G))},$$

which is the claim. The denominator is strictly positive whenever some $t_j > 0$, since $\alpha\, v_0^\top D^{-1} v_0 = \alpha\sum_j (v_0)_j^2/(t_j+\rho\lambda_2(K_G)) \le \alpha\,\|v_0\|^2/(\rho\lambda_2(K_G)) = 1$, with equality only when all $t_j = 0$. □

Corollary 18 (Asymptotic floor) *If $t_j \ge \rho\lambda_2(K_G)\,n\,(v_0)_j^2$ for all j, then $\varepsilon_T \le 1/2$ and*

$$t_{\mathrm{eff},i}(T) \;\ge\; \big(t_i + \rho\lambda_2(K_G)\big)\cdot \frac{1}{1 + \frac{2\rho\lambda_2(K_G)\,(v_0)_i^2}{t_i+\rho\lambda_2(K_G)}}.$$

For the combinatorial Laplacian, $(v_0)_i^2 = 1/n$ uniformly and the right-hand correction is $O(1/n)$, recovering the standard uniform-floor bound. For the normalized Laplacian, $(v_0)_i^2 = \deg_G(i)/(2|E_G|)$ varies with degree: high-degree (hub) arms have larger $(v_0)_i^2$ and a weaker effective-count floor, while low-degree (leaf) arms have $(v_0)_i^2 \to 0$ and a tighter floor. This degree-dependent profile is the analytical counterpart of the BA experiments in Appendix H.2.

Proposition 19 (Spectral upper bound) *For any PSD kernel K_G with largest eigenvalue $\lambda_n(K_G)$ and any sampling counts $\{t_j\}$,*

$$t_{\mathrm{eff},i}(T) \;\le\; t_i + \rho\lambda_n(K_G).$$

Proof: $K_G \preceq \lambda_n(K_G)\, I$, so $V_T \preceq N_T + \rho\lambda_n(K_G)\, I$, hence $V_T^{-1} \succeq \big(N_T + \rho\lambda_n(K_G) I\big)^{-1}$, and $[V_T^{-1}]_{ii} \geq 1/(t_i + \rho\lambda_n(K_G))$. □

Remark 20 (Kernel-explicit hardness) *Combining Propositions 17 and 19, the per-arm threshold $t_{\mathrm{eff},i}(T) \geq 196\, C(T)L_2(T)/\Delta_{i,c}^2$ from the elimination lemma is achievable for direct-pull budgets sandwiched as*

$$\frac{196\, C(T)L_2(T)}{\Delta_{i,c}^2} - \rho\lambda_n(K_G) \;\leq\; t_i^\star \;\leq\; \frac{196\, C(T)L_2(T)}{\Delta_{i,c}^2} - \rho\lambda_2(K_G)\big(1 - (v_0)_i^2\big).$$

The condition number $\kappa(K_G) := \lambda_n(K_G)/\lambda_2(K_G)$ measures the range between best-case (well-aligned) and worst-case (poorly-aligned) information sharing across arms. The kernel-modified hardness obeys

$$H_{\mathrm{graph}}^{(K_G)} \;\leq\; \sum_{i \neq a^\star} \frac{1}{\big(t_i + \rho\lambda_2(K_G)(1 - (v_0)_i^2)\big)\, \Delta_{i,c}^2},$$

and the hardness improves linearly with the spectral gap. For the combinatorial Laplacian, $(v_0)_i^2 = 1/n$ for all i and the correction $1 - (v_0)_i^2$ is uniform. For the normalized Laplacian on a non-regular graph, $(v_0)_i^2 = \deg_G(i)/(2|E_G|)$ varies with the arm and the per-arm hardness is degree-dependent.

Remark 21 (Combinatorial vs. normalized Laplacian) *For the combinatorial Laplacian L_G, Cauchy interlacing gives $\lambda_n(L_G) \leq 2d_{\max}$, while $\lambda_2(L_G)$ (the algebraic connectivity) can be as small as $\Theta(1/n)$ on poorly-connected graphs, yielding $\kappa(L_G) = \Omega(nd_{\max})$ in the worst case. For the normalized Laplacian $\mathcal{L}_G = I - D_G^{-1/2} A_G D_G^{-1/2}$, $\lambda_n(\mathcal{L}_G) \leq 2$ regardless of degree heterogeneity and $\lambda_2(\mathcal{L}_G) \geq \lambda_2(L_G)/d_{\max}$ [15], so $\kappa(\mathcal{L}_G) \leq 2d_{\max}/\lambda_2(L_G)$. For example, on a star graph with n leaves, $\kappa(L_G) \asymp n$ while $\kappa(\mathcal{L}_G) = O(1)$, and the gap predicted by Remark 20 corresponds to a factor-of-n improvement in $H_{\mathrm{graph}}^{(K_G)}$. This is the regime in which the normalized Laplacian $\mathcal{L}_G$ is the recommended choice of kernel.*

Sample complexity (analogue of Theorem 5).

Proof: By the elimination lemma, it suffices for each suboptimal arm i that

$$t_{\mathrm{eff},i}^{(\mathrm{norm})}(t) \geq \frac{196\, C(t)L_2(t)}{\Delta_{i,c}^2}.$$

Since $t_{\mathrm{eff},i}^{(\mathrm{norm})}(t) = [(V_t)^{-1}]_{ii}^{-1}$ with $V_t = \sum_{s=1}^t e_{\pi_s} e_{\pi_s}^\top + \rho K_G$, Proposition 17 (combined with the graph-influence argument of Theorem 4, both of which use only the PSD property of the regularizer) gives

$$t_{\mathrm{eff},i}^{(\mathrm{norm})}(t) \geq t_i(t) + \frac{1}{2}\left\lfloor \min\left\{ \rho\, \mathcal{J}^{(\mathrm{norm})}(i, G),\; \sum_{j \in C(i)} t_j(t) \right\} \right\rfloor,$$

where $\mathcal{J}^{(\mathrm{norm})}(i, G)$ is the influence factor computed using the kernel K_G (equivalently, using the resistance distances induced by K_G), and is itself lower-bounded by $\rho\lambda_2(K_G)(1 - 1/n)$ via Proposition 17. The proof of Theorem 4 uses only the PSD property of the regularizer and the resulting matrix-inversion structure, both of which hold for any PSD kernel.

Partitioning the suboptimal arms into competitive and non-competitive sets proceeds identically:

- *Competitive arms* ($i \in \mathcal{H}^{(\mathrm{norm})} \setminus \{a^\star\}$): those for which the kernel influence factor does not suffice to satisfy the threshold, requiring direct pulls at a cost of $\sum_{i \in \mathcal{H}^{(\mathrm{norm})} \setminus \{a^\star\}} C(t) L_2(t) / \Delta_{i,c}^2$.
- *Non-competitive arms* ($i \in \mathcal{N}^{(\mathrm{norm})}$): eliminated quickly because their kernel-effective counts are augmented through the graph, at a cost of $\max_{i \in \mathcal{N}^{(\mathrm{norm})}} C(t) L_2(t) / \Delta_{i,c}^2$.

Combining both contributions yields the self-referential equation

$$T \le c\, C(T) L_2(T) \left(\sum_{i \in \mathcal{H}^{(\mathrm{norm})} \setminus \{a^\star\}} \frac{1}{\Delta_{i,c}^2} + \max_{i \in \mathcal{N}^{(\mathrm{norm})}} \frac{1}{\Delta_{i,c}^2} \right) = c\, C(T) L_2(T) \cdot H_{\mathrm{graph}}^{(\mathrm{norm})},$$

where $H_{\mathrm{graph}}^{(\mathrm{norm})}$ denotes the graph hardness computed with kernel-effective counts. Solving the self-referential logarithm as in the proof of Theorem 5 gives

$$T = O\left(H_{\mathrm{graph}}^{(\mathrm{norm})} \left(\log \frac{1}{\delta} + \log\Big(K H_{\mathrm{graph}}^{(\mathrm{norm})} \Big) \right) \frac{\log(1/\delta)}{\log(1/q)} + H_{\mathrm{graph}}^{(\mathrm{norm})} \frac{\log^2(K H_{\mathrm{graph}}^{(\mathrm{norm})})}{\log(1/q)} \right).$$

Setting $q = \delta$ yields $O\Big(H_{\mathrm{graph}}^{(\mathrm{norm})} \log(1/\delta) + H_{\mathrm{graph}}^{(\mathrm{norm})} \log^2(K H_{\mathrm{graph}}^{(\mathrm{norm})}) \Big)$. □

G Proofs for Misspecified Smoothness (Conservative Budget)

Recall that in this section the algorithm is run with a nominal smoothness parameter ε, whereas the true reward vector satisfies

$$\|\boldsymbol{\mu}\|_G \le \varepsilon_{\mathrm{true}}, \qquad \varepsilon_{\mathrm{true}} \ge \varepsilon.$$

Thus the misspecification is that the smoothness budget used in the confidence radius is too small. Define

$$\Gamma_{\mathrm{mis}} := \varepsilon_{\mathrm{true}} - \varepsilon \ge 0.$$

The good-event analysis below establishes guarantees for the algorithm *as run* with the nominal ε, provided the Thompson sampling variance $C(t)$ is set with respect to a known upper bound $\overline{\varepsilon} \ge \varepsilon_{\mathrm{true}}$ on the smoothness. Concretely, the algorithm draws samples with variance $C^{\mathrm{mis}}(t)/t_{\mathrm{eff},i}(t)$ where

$$C^{\mathrm{mis}}(t) \;:=\; \frac{2\big(\sigma_0 \sqrt{L_1(t)} + \sqrt{\rho}\,\overline{\varepsilon}\big)^2}{\phi^2(q)};$$

in the well-specified case $\overline{\varepsilon} = \varepsilon = \varepsilon_{\mathrm{true}}$ and $C^{\mathrm{mis}}(t)$ reduces to the $C(t)$ of Section 3.1. The bias-radius parameter $\overline{\varepsilon}$ used in $C(t)$ must be a valid upper bound on the smoothness for the survival argument of G_2 to go through. Using a strict under-estimate (i.e. $\overline{\varepsilon} < \varepsilon_{\mathrm{true}}$) breaks the union bound. Whenever such an upper bound is available, the analysis that follows shows the elimination structure of Theorem 7 holds with $\varepsilon_{\mathrm{true}}$ in place of ε in the hardness expression.

Lemma 22 (Good event for the under-specified model) *Suppose the algorithm is run with nominal smoothness budget ε while the true reward vector satisfies $\|\boldsymbol{\mu}\|_G \le \varepsilon_{\mathrm{true}}$ with $\varepsilon_{\mathrm{true}} \ge \varepsilon$. Define the misspecified confidence radius*

$$b_i^{\mathrm{mis}}(t) := \sqrt{\frac{1}{t_{\mathrm{eff},i}(t)}} \left(2\sigma \sqrt{14 \log\Big(\tfrac{2 w_i(\boldsymbol{\pi}_t)}{\delta} \Big)} + \sqrt{\rho}\, \varepsilon_{\mathrm{true}} \right),$$

the events

$$G_0^{\mathrm{mis}} := \big\{\forall t \geq 1,\ \forall i \in [K] :\ |\widehat{\mu}_i(t) - \mu_i| \leq b_i^{\mathrm{mis}}(t)\big\},$$
$$G_1 := \big\{\forall t, m, i :\ |\theta_i^m(t) - \widehat{\mu}_i(t)| \leq \sqrt{2C(t)L_2(t)/t_{\mathrm{eff},i}(t)}\big\},$$
$$G_2^{\mathrm{mis}} := \big\{\forall t \geq 1,\ \forall i \neq j,\ \exists m \in [M(\delta, q, t)]\ \textit{s.t.}\ \theta_i^m(t) - \theta_j^m(t) \geq \mu_i - \mu_j\big\}.$$

Then $\Pr\big(G_0^{\mathrm{mis}} \cap G_1 \cap G_2^{\mathrm{mis}}\big) \geq 1 - \delta$.

Proof of Lemma 22. The proof follows the same structure as the proof of the corresponding good event in the well-specified model. The only difference is that the estimator bias term now involves the *true* smoothness level $\varepsilon_{\mathrm{true}}$.

For each arm i and time t, define

$$b_i^{\mathrm{mis}}(t) := \sqrt{\frac{1}{t_{\mathrm{eff},i}(t)}}\left(2\sigma\sqrt{14\log\left(\frac{2w_i(\boldsymbol{\pi}_T)}{\delta}\right)} + \sqrt{\rho}\,\varepsilon_{\mathrm{true}}\right).$$

Let G_0^{mis} denote the event that

$$|\widehat{\mu}_i(t) - \mu_i| \leq b_i^{\mathrm{mis}}(t) \qquad \text{for all } i \in [K],\ t \geq 1.$$

Since the proof of the estimator concentration inequality depends only on the upper bound on $\|\boldsymbol{\mu}\|_G$, the same argument used in the well-specified case yields

$$\Pr\big(G_0^{\mathrm{mis}}\big) \geq 1 - \frac{\delta}{3},$$

provided the constants are chosen exactly as in the original proof, except with $\varepsilon_{\mathrm{true}}$ replacing ε in the bias term. In particular, the only place where smoothness enters is through the deterministic bound

$$|\langle e_i, \rho V_t^{-1} L_G \boldsymbol{\mu}\rangle| \leq \rho\sqrt{[V_t^{-1}]_{ii}}\,\sqrt{\langle L_G\boldsymbol{\mu}, V_t^{-1}L_G\boldsymbol{\mu}\rangle} \leq \sqrt{\rho}\,\sqrt{[V_t^{-1}]_{ii}}\,\|\boldsymbol{\mu}\|_G \leq \sqrt{\rho}\,\sqrt{[V_t^{-1}]_{ii}}\,\varepsilon_{\mathrm{true}},$$

which gives the additional additive term $\sqrt{\rho}\,\varepsilon_{\mathrm{true}}/\sqrt{t_{\mathrm{eff},i}(t)}$ in the radius. The factor of $\sqrt{\rho}$ rather than ρ comes from $V_t \succeq \rho L_G \Rightarrow V_t^{-1} \preceq \frac{1}{\rho}L_G^{\dagger}$ on the range of L_G, so $\langle L_G\boldsymbol{\mu}, V_t^{-1}L_G\boldsymbol{\mu}\rangle \leq \frac{1}{\rho}\|\boldsymbol{\mu}\|_G^2$.

Next, let G_1 be the same Thompson-sample concentration event as in the well-specified proof. Its proof is unchanged, since it depends only on the Gaussian sampling step conditioned on the empirical estimator and not on the graph smoothness level. Therefore

$$\Pr(G_1) \geq 1 - \frac{\delta}{3}$$

for the same choice of constants as before.

It remains to control the event G_2^{mis}, namely that at each round there exists at least one Thompson sample whose pairwise comparisons preserve the true arm ordering with probability at least q. Fix a time t, two arms i, i', and condition on the event G_0^{mis}. By definition of the event G_0^{mis}, we have

$$\widehat{\mu}_i(t) \geq \mu_i - b_i^{\mathrm{mis}}(t), \qquad \widehat{\mu}_{i'}(t) \leq \mu_{i'} + b_{i'}^{\mathrm{mis}}(t).$$

Hence

$$\widehat{\mu}_i(t) - \widehat{\mu}_{i'}(t) \geq \mu_i - \mu_{i'} - b_i^{\mathrm{mis}}(t) - b_{i'}^{\mathrm{mis}}(t).$$

Now, conditioned on the filtration up to time t, the Thompson difference

$$\theta_i^k(t) - \theta_{i'}^k(t)$$

is Gaussian with mean

$$\widehat{\mu}_i(t) - \widehat{\mu}_{i'}(t)$$

and variance

$$C_{\mathrm{mis}}(t)\left(\frac{1}{t_{\mathrm{eff},i}(t)} + \frac{1}{t_{\mathrm{eff},i'}(t)}\right).$$

Therefore,

$$\begin{aligned}
&\Pr\left[\theta_i^k(t) - \theta_{i'}^k(t) \geq \mu_i - \mu_{i'} \;\middle|\; G_0^{\mathrm{mis}}\right] \\
&\quad = \Phi\left(\mu_i - \mu_{i'}, \widehat{\mu}_i(t) - \widehat{\mu}_{i'}(t), C_{\mathrm{mis}}(t)\left(\frac{1}{t_{\mathrm{eff},i}(t)} + \frac{1}{t_{\mathrm{eff},i'}(t)}\right)\right) \\
&\quad \geq \Phi\left(b_i^{\mathrm{mis}}(t) + b_{i'}^{\mathrm{mis}}(t), 0, C_{\mathrm{mis}}(t)\left(\frac{1}{t_{\mathrm{eff},i}(t)} + \frac{1}{t_{\mathrm{eff},i'}(t)}\right)\right).
\end{aligned}$$

Thus, if $C_{\mathrm{mis}}(t)$ is chosen large enough so that

$$\Phi\left(\frac{b_i^{\mathrm{mis}}(t) + b_{i'}^{\mathrm{mis}}(t)}{\sqrt{C_{\mathrm{mis}}(t)\left(\frac{1}{t_{\mathrm{eff},i}(t)} + \frac{1}{t_{\mathrm{eff},i'}(t)}\right)}}, 0, 1\right) \geq q$$

for every pair (i, i'), then each Thompson draw has probability at least q of certifying the desired comparison. Taking

$$M(\delta, q, t)$$

independent Thompson draws at round t, the probability that all such draws fail is at most $(1-q)^{M(\delta,q,t)}$, and by the same choice of $M(\delta, q, t)$ and the same union bound as in the well-specified analysis, we obtain

$$\Pr\big(G_2^{\mathrm{mis}} \mid G_0^{\mathrm{mis}}\big) \geq 1 - \frac{\delta}{3}.$$

Combining the bounds for G_0^{mis}, G_1, and G_2^{mis}, and applying a union bound, yields

$$\Pr\big(G_0^{\mathrm{mis}} \cap G_1 \cap G_2^{\mathrm{mis}}\big) \geq 1 - \delta.$$

This proves the lemma. □

Correctness under the misspecified event. The correctness argument is the same as in the well-specified case once we work on the good event

$$G_0^{\mathrm{mis}} \cap G_1 \cap G_2^{\mathrm{mis}}.$$

Under G_2^{mis}, at every round there exists a Thompson draw that separates the empirical best arm from any truly suboptimal arm. A suboptimal arm therefore cannot remain consistent with all Thompson samples once the elimination rule is invoked. Hence the algorithm cannot terminate and output a suboptimal arm. Termination occurs only when all remaining arms are consistent with the Thompson comparisons, so the returned arm is optimal. □

Sample complexity under under-specification. We now show how the stopping-time bound changes. The proof is identical in structure to the well-specified case, but every occurrence of the nominal confidence scale is replaced by the misspecification-adjusted one.

Fix a suboptimal arm i. Under the event $G_0^{\mathrm{mis}} \cap G_1 \cap G_2^{\mathrm{mis}}$, arm i is eliminated as soon as its confidence width is small enough relative to the gap $\Delta_{i,c}$. There exists an absolute constant $c > 0$ such that it suffices to have

$$\sqrt{\frac{C_{\mathrm{mis}}(t)L(t)}{t_{\mathrm{eff},i}(t)}} \le \frac{\Delta_{i,c}}{c},$$

or equivalently,

$$t_{\mathrm{eff},i}(t) \ge c\,\frac{C_{\mathrm{mis}}(t)L(t)}{\Delta_{i,c}^2}.$$

This is the same elimination condition as before, with the variance scale $C(t)$ replaced by $C_{\mathrm{mis}}(t)$.

At this point the proof proceeds as in the main sample-complexity argument. The lower bound on the effective sample size,

$$t_{\mathrm{eff},i}(t),$$

depends only on the graph geometry and the sampling policy and is unchanged by misspecification of the smoothness budget. The proof of the lower bound on $t_{\mathrm{eff},i}(t)$ applies without modification, since it does not use the value of ε. The same holds in the GRUB-style effective-sample analysis, where the lower bound on $t_{\mathrm{eff},i}$ depends on graph structure and pull counts rather than the smoothness radius.

The competitive/non-competitive decomposition therefore retains the same structure. The only change is that the threshold defining whether an arm is competitive is now formed using the larger confidence scale

$$\frac{2}{\sqrt{\rho\,\mathcal{J}(i,G)}}\left(\sigma_0\sqrt{L_1(t)} + \sqrt{\rho}\,\varepsilon_{\mathrm{true}}\right)$$

instead of the nominal quantity based on ε. Since $\varepsilon_{\mathrm{true}} \ge \varepsilon$, the competitive set can only increase, and the non-competitive set can only decrease.

Therefore, after replacing the original hardness parameter by the enlarged one

$$H_{\mathrm{mis}} \asymp \sum_{i \in \mathcal{H}_{\mathrm{mis}} \setminus \{a^\star\}} \frac{1}{\Delta_{i,c}^2} + \max_{i \in \mathcal{N}_{\mathrm{mis}}} \frac{1}{\Delta_{i,c}^2},$$

the remainder of the stopping-time proof is unchanged and yields the stated bound in Theorem 7. □

On under-estimation. The analysis above treats the case $\bar{\varepsilon} \ge \varepsilon_{\mathrm{true}}$, in which the Thompson variance scale $C^{\mathrm{mis}}(t)$ is conservative. The reverse case $\bar{\varepsilon} < \varepsilon_{\mathrm{true}}$ is qualitatively different: the survival argument for G_2 requires the Thompson scale to dominate the worst-case bias, and a strict under-estimate breaks this calibration so the union bound for G_2 no longer holds. In practice, an upper bound $\bar{\varepsilon}$ can be obtained from a residual-based estimate $\widehat{\varepsilon}^2 := \langle \widehat{\boldsymbol{\mu}}, L_G \widehat{\boldsymbol{\mu}} \rangle$ at any round, inflated by a high-probability correction term derived from the same concentration inequality used for G_0. Detection of under-estimation (when $\widehat{\varepsilon}$ exceeds the nominal $\bar{\varepsilon}$) can then trigger a doubling of the budget. We do not pursue this extension here.

H Experimental Details

This appendix collects the protocol, hyperparameters, and additional experiments referenced in Section 4.

H.1 Implementation, baselines, and reproducibility

Algorithms. For the Laplacian-regularized estimators (TS-Explore and GRUB), each round forms the regularized design matrix $V_t = \sum_{s \le t} e_{\pi_s} e_{\pi_s}^\top + \rho L_G + \rho_{\text{diag}} I$, where $\rho_{\text{diag}} = \max(10^{-4}, 10^{-6}\rho)$ guarantees invertibility for arbitrarily large ρ. We compute V_t^{-1} by Cholesky factorization. The agreement stopping rule draws $M(\delta, q, t)$ Gaussian Thompson copies per arm at each round with sampling variance $C(\delta, q, t)/t_{\text{eff},i}(t)$, using M and C as defined in Section 3.1. The pull rule selects the less-sampled arm in the disagreement pair $\{\widehat{i}_t, \widetilde{i}_t\}$, with arbitrary tie-breaking. Initialization pulls one arm from each connected component of G to ensure V_t is full-rank from the start. For the Erdős–Rényi sweep this means the number of initial pulls varies with the realized component count at small p.

The graph-feedback algorithms (TS-Explore-GF, UCB+cover, TS+width, UCB-N) use the empirical mean $\widehat{\mu}_i(t) = R_i^{\text{fb}}(t)/N_i^{\text{fb}}(t)$ over side observations rather than a Laplacian-regularized estimator. They differ along two design axes. The first is the stopping rule, which is either agreement-based (TS) or UCB–LCB elimination (UCB). The second is the pull rule, which is either the candidate-pair cover $\arg\max_a |N_G^+(a) \cap \{\widehat{i}_t, \widetilde{i}_t\}|$ (with ties broken by smaller N_a^{fb}) or argmax confidence width over the remaining set. Initialization is a greedy dominating-set covering of G so every arm is observed at least once.

Stopping rules and confidence. The TS variants use the agreement-based stopping rule of Algorithm 1. GRUB and UCB-N use UCB–LCB elimination. KL-LUCB uses a sub-Gaussian Hoeffding-style radius with the LUCB sampling rule of [27]. All algorithms target the same confidence $\delta = 10^{-3}$. The agreement quantile is fixed to $q = 0.1$ throughout, which is the loosest setting compatible with the analysis. Concretely:

- KL-LUCB uses the anytime confidence width $c_i(t) = \sigma\sqrt{2\,\log(Kt^2/\delta)/N_i(t)}$ on the empirical mean. It stops when the lower bound on the empirical leader exceeds the upper bound on its closest challenger.
- UCB-N uses the side-observation confidence width $\sigma\sqrt{2\,L_1(t)/N_i^{\text{fb}}(t)}$ with $L_1(t) = \log\big(12K^2t^2/\delta\big)$, matching the variance term in the radius of TS-Explore-GF so that the comparison isolates the stopping and pull rules. The pull rule is the pure-exploration adaptation of UCB-N [11]: pull the arm in the candidate set with the largest confidence width. We deviate from the standard arg max UCB rule, which fixes on the empirical leader and fails to discriminate sparsely-observed arms in fixed-confidence best-arm identification.

GRUB uses the radius $\beta_i(t) = 2\sigma\sqrt{14\,\log(2Kt/\delta)} + \rho\varepsilon$ of [47] on the Laplacian-regularized estimator, with the same Laplacian regularization ρ as TS-Explore on a per-experiment basis ($\rho = 100$ on the synthetic instance, swept jointly with TS-Explore on MovieLens). For the arm-selection rule we use the marginal-variance variant of GRUB described in [47], which pulls the candidate arm with the largest current marginal variance, that is, $\arg\max_i [V_t^{-1}]_{ii}$ over the candidate set.

Replication, budget, and seeds. Main-body cells (Section 4) are replicated over 20 seeds, and appendix cells over 5 to 10 seeds depending on per-cell wall-clock. Per seed, the random state controls reward draws, Thompson samples, the random-graph realization where applicable (Erdős–Rényi, stochastic block model, Barabási–Albert), and tie-breaking. We report the per-cell median and the 25–75 interquartile range. The per-run budget is 10^7 pulls. The only configuration that hit this cap is GRUB at $K = 200$ on the synthetic chain, where all 20 seeds reach the cap. See the GRUB-scaling discussion in Section 4. Outside this case, every completed run returned the optimal arm.

Datasets and graph construction. For MovieLens-100K [24], we restrict to the K movies with the largest number of ratings. The arm reward distributions ν_i are the empirical rating distributions of these movies, sampled with replacement, and the ground-truth means μ_i are the corresponding empirical rating means. The item-item similarity graph is constructed from the user-rating matrix using the standard adjusted-cosine similarity, which is the cosine similarity between rating vectors after subtracting each user's mean rating. We then sparsify the similarity graph by retaining each pair (i, j) as an edge only if j is among the top-k neighbors of i *and* i is among the top-k neighbors of j (mutual top-k). For the stochastic block model and Barabási–Albert experiments, exact generative parameters are given in the relevant subsections below.

Compute. All experiments were run on a single workstation with a 24-core Intel CPU, 48 GB of RAM, and Ubuntu 24.04. No GPU was used. The full reported sweep, including all main-body and appendix cells, can be reproduced on a comparable CPU-only machine in under 100 CPU-hours. Wall-clock is dominated by the high-ρ MovieLens runs and the small-ε misspecification runs, where small effective sample sizes per round translate into long stopping times. The most expensive single cell is the $K = 200$ GRUB cell on the synthetic chain, which all 20 seeds drive into the per-run 10^7-pull cap.

Code and data availability. Code and figure-regeneration scripts are provided in the anonymous supplementary materials accompanying this submission. The MovieLens-100K dataset is publicly available at `grouplens.org/datasets/movielens/100k/` and is auto-downloaded by the experiment scripts on first run.

H.2 Normalized vs combinatorial Laplacian

This experiment tests Theorem 14 and the spectral sandwich of Proposition 15 on a graph for which the two propositions predict the largest gap between the combinatorial and the normalized Laplacian: a degree-heterogeneous Barabási–Albert graph. We use $n = 50$ vertices and preferential-attachment parameter $m = 2$, with the optimal arm fixed at the highest-degree hub (ties broken by the smallest vertex index). Across the 10 seeds, the realized hub degree has median 15 (range $[11, 18]$) and the realized leaf degree has median 2 (range $[1, 2]$). We assign a uniform suboptimality gap $\Delta = 0.3$ to every other arm and sweep ρ over a logarithmic grid, running TS-Explore with two graph kernels: the combinatorial Laplacian L_G and the normalized Laplacian $K_G = I - D_G^{-1/2} A_G D_G^{-1/2}$.

Figure 3 shows both kernels matching Basic TS at small ρ and degrading at large ρ, with empirical optima on our ρ-grid at $\rho^{\text{emp}}_{L_G} = 10$ and $\rho^{\text{emp}}_{K_G} = 300$. The combinatorial Laplacian degrades sharply once ρ exceeds $\rho^{\text{emp}}_{L_G}$, while the normalized Laplacian remains near its minimum past $\rho^{\text{emp}}_{K_G}$. At each kernel's empirical optimum, the median stopping-time ratio is 1.12 in favor of K_G, with a 95% bootstrap confidence interval of $[1.05, 1.19]$ over 10 seeds. The absolute minimum of TS-L_G at $\rho = 10$ is approximately 203,510 pulls, which is above the no-graph Basic TS baseline of 201,772. So L_G does not improve over Basic TS at any ρ on this instance, while K_G does.

This is the qualitative pattern that Proposition 15 predicts. The condition number $\kappa(K_G) = \lambda_n/\lambda_2$ controls the uniformity of effective counts across arms, and the normalized Laplacian satisfies $\lambda_n(\mathcal{L}_G) \leq 2$ universally while the combinatorial Laplacian's largest eigenvalue grows with the maximum degree of G. On a graph with strong degree heterogeneity, the combinatorial Laplacian therefore allocates effective counts unevenly enough to wipe out the graph-bonus advantage. The normalized Laplacian recovers a regime in which the kernel formulation provides a measurable improvement over the unstructured baseline.

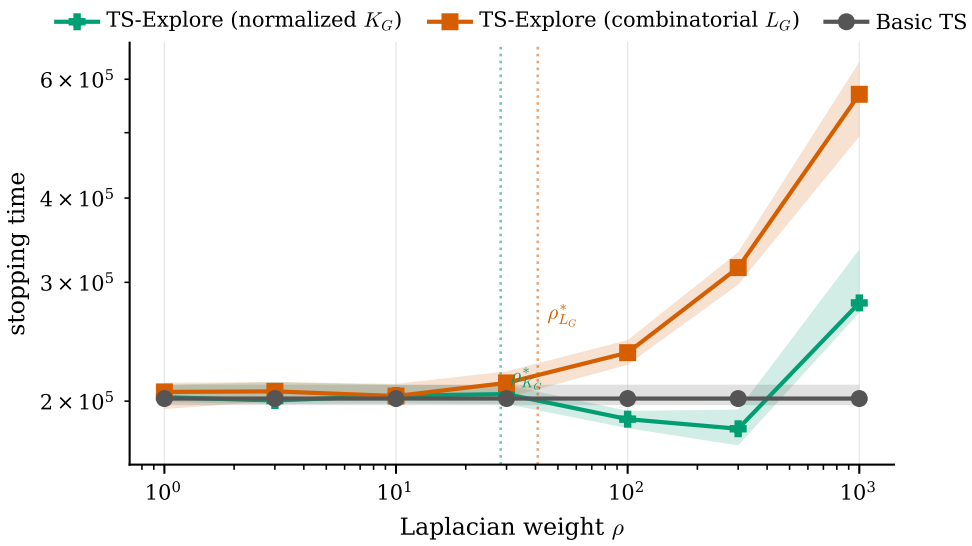


Figure 3: Normalized vs combinatorial Laplacian on a Barabási–Albert graph ($n = 50$, $m = 2$, uniform $\Delta = 0.3$, hub-optimal). Both kernels match Basic TS at small ρ and degrade at large ρ. Vertical dotted lines mark the empirical optimum ρ^{emp} for each kernel. Median and 25–75 IQR over 10 seeds.

H.3 Smoothness asymptotics on a stochastic block model

This experiment tests the smoothness-asymptotics result of Corollary 8, which predicts that the stopping time of TS-Explore run with $\rho \geq \rho_{\text{var}}(\varepsilon)$ should decrease as $\varepsilon \to 0$ on instances where every suboptimal arm has a positive gap. We use a connected stochastic block model with $K = 31$ arms. The instance has 5 challenger blocks of size 6 (giving 30 challenger arms) plus an isolated best arm at index 0, intra-block edge probability 0.9, and four bridge edges between consecutive cluster representatives at indices 1→7→13→19→25. The bridges ensure that the realized graph is one connected component of size 31, so the asymptotic limit $\max_i \Delta_i^{-2} \log(1/\delta)$ from the corollary applies. Cluster means are linearly spaced as $[0.9, 0.7, 0.5, 0.3, 0.1]$ across the five blocks, the best arm has $\mu_{a^\star} = 1$, and these settings give $\sum_{i \neq a^\star} \Delta_i^{-2} = 710.32$ and $\max_{i \neq a^\star} \Delta_i^{-2} = 100$. We sweep ε over the logarithmic grid $\{10^{-k/2}\}_{k=0,\ldots,6}$ and run three algorithms over 5 seeds each: TS-Explore with $\rho \geq \rho_{\text{var}}(\varepsilon)$, TS-Explore at $\rho = 1$, and Basic TS.

Table 2 reports per-ε medians. The recommended-ρ curve decreases monotonically from $1.37 \cdot 10^6$ at $\varepsilon = 1$ to a minimum of $1.12 \cdot 10^5$ at $\varepsilon = 10^{-2.5}$, with a small increase to $1.31 \cdot 10^5$ at $\varepsilon = 10^{-3}$. At its minimum, TS-Explore-ρ_{var} is 11.3 times faster than the un-tuned baselines, and the un-tuned baselines stay constant across the sweep, as expected. Reaching the asymptotic floor $\max_i \Delta_i^{-2} \log(1/\delta) \approx 690$ would require ε below the smallest critical level $\varepsilon_i^\star$ from Eq. (6), which is outside the swept range. The realized minimum at $\varepsilon = 10^{-2.5}$ is approximately 162 times the floor.

These observations match the qualitative shape predicted by Corollary 8. Smoothness-aware regularization tracks the smoothness budget, so the recommended-ρ curve descends as $\varepsilon \to 0$. The un-tuned baselines have no mechanism to exploit small ε, so they remain at the no-graph rate. The experiment confirms that the asymptotic improvement promised by the corollary is realizable on a non-trivial graph instance, even when the swept range does not reach all the way to the analytical floor.

Table 2: Smoothness asymptotics on a stochastic block model. Median stopping time over 5 seeds at each ε for the three algorithms. The competitive-set size $|\mathcal{H}_\varepsilon|$ undergoes phase transitions ($30 \to 18 \to 12 \to 6$) as ε shrinks past successive critical levels $\varepsilon_i^\star$ from Eq. (6), illustrating Corollary 8. TS-Explore-$\rho_{\mathrm{var}}(\varepsilon)$ minimises at $\varepsilon = 10^{-2.5}$, where it is roughly an order of magnitude faster than the un-tuned baselines.

ε	1	$10^{-0.5}$	10^{-1}	$10^{-1.5}$	10^{-2}	$10^{-2.5}$	10^{-3}
H_ε	710	710	710	710	692	670	611
$\lvert\mathcal{H}_\varepsilon\rvert$	30	30	30	30	18	12	6
TS-Explore, $\rho_{\mathrm{var}}(\varepsilon)$	1,372,441	1,358,820	1,339,398	964,078	295,836	**112,409**	130,828
TS-Explore, $\rho = 1$	1,271,404	1,271,404	1,271,404	1,271,404	1,271,404	1,271,404	1,271,404
Basic TS	1,271,375	1,271,375	1,271,375	1,271,375	1,271,375	1,271,375	1,271,375